Population structure and adaptation in fishes: Insights from clupeid and salmonid species

Limborg, Morten

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Population structure and adaptation in fishes

Insights from clupeid and salmonid species

PhD thesis by

Morten Tønsberg Limborg

December 2011

Technical University of Denmark

National Institute of Aquatic Resources

Section for Population Ecology and Genetics

Cover photos: Top: Finn Sivebæk, Bottom: *Unknown*

Main supervisor:

Dorte Bekkevold
National Institute of Aquatic Resources
Technical University of Denmark

Co-supervisors:

Einar Eg Nielsen
National Institute of Aquatic Resources
Technical University of Denmark

Brian MacKenzie
National Institute of Aquatic Resources
Technical University of Denmark

Michael Møller Hansen
Department of Bioscience
University of Aarhus

Contents

Preface

This thesis was submitted to the DTU Aqua PhD School of the Technical University of Denmark in December 2011 for partial fulfilment of obtaining the PhD degree in biology. The work conducted over the past three years during this PhD programme has mainly taken place at DTU Aqua in Silkeborg including a four months external research visit at the School of Aquatic and Fishery Sciences at the University of Washington, Seattle, USA in the summer 2010. Funding for the PhD programme was provided by DTU Aqua and the European Commission through the projects UNCOVER and RECLAIM. I am further grateful for two travel grants from the Otto Mønsted foundation for going to Seattle twice.

The overall objective of this PhD project was to enhance our understanding of the evolutionary processes shaping population structure and enabling adaptations to local environments in fishes. Three species representing two eco-types have been studied during this work; the marine small pelagic fishes Atlantic herring (*Clupea harengus*) and European sprat (*Sprattus sprattus*) as well as the salmonid *Oncorhynchus mykiss*. We used mtDNA and microsatellite markers to describe the demographic history and contemporary population structure of sprat (*chapters 2-4*). Then, considering the benefits of applying large marker panels for detecting signatures of natural selection (*chapter 5*), we developed a new panel of SNP markers for herring (*chapter 6*). This panel was used to study the occurrence and extent of local adaptation throughout the northeastern Atlantic distribution of herring (*chapter 7*) and a similar approach was applied to detect spatial signatures of adaptively important candidate genes in *O. mykiss* (*chapter 8*). In *chapter 9* I conclude with a perspective of where I believe future efforts within the field are expected to significantly increase our perception of evolutionary processes in the wild.

Here I give a brief guide through the content of the general introduction (*chapter 1*) that serves to state the background and objectives of my PhD as well as introducing relevant theory and methodological issues of relevance to this work. First, I state the importance of studying and understanding evolutionary processes in the wild, and this is followed by a short presentation of the European 7th Framework Programme project FishPopTrace (FPT), which has had a leading role during my PhD. FishPopTrace, a three year research programme, has been running over the course of my PhD, and a large part of my research (*chapters 5-7*) relates to work carried out in collaboration with this consortium. This is followed by a short paragraph where I place genetic studies in a larger biological framework needed to fully comprehend how genetic variation of

organisms is affected by external parameters like the surrounding environment. After this, I give a more technical presentation of relevant molecular markers for assessing genetic population structure and scales of local adaptation. This is followed by a description of different analytical methods for detecting signatures of local adaptation with a strong focus on three approaches (i.e. genome scans, candidate genes, and landscape genetics), which I have used extensively in my efforts for detecting local adaptation in Atlantic herring (*chapter 7*) and *O. mykiss* (*chapter 8*). This is followed by a critical discussion of some analytical limitations that need careful consideration for interpreting results indicating local adaptation. Hereafter, I introduce my three study species, with a focus on the ecological characteristics making them attractive models for addressing questions of population structure and local adaptation in fish. Lastly, I end chapter 1 with a brief background of the projects and motivations for conducting the different studies making up this thesis.

Where relevant, the independent manuscripts making up this thesis will be referred to as *chapters 2-8* throughout *chapter 1* and *9*. I reckon that vast amounts of relevant studies for the general topics discussed in the following chapters exist, and my citations suffer an inevitable bias towards the more fishy side of the literature. This does by no means reflect a subjective perception that ignored literature is inferior, but simply because I am more familiar with the cited examples.

The final outcome of this PhD has only been made possible from invaluable help and collaboration from numerous people in all sorts of ways, and I apologise to those I may forget here – thank you. I have been blessed with the opportunity to work with an endless number of bright collegial biologists from all over Europe including the entire FishPopTrace consortium, especially Sarah, Ilaria, Babbucci, Alessia, Greg, Martin and Gary. Good times were also spent with fellow students from the DTU AQUA PhD school and a lot of people at DTU Aqua in Silkeborg. I am grateful for having been part of a team full of really bright population geneticists (Jakob, Thomas, Michael, and Einar) and invaluable and talented lab technicians (Karen-Lise, Dorte, and especially Noor for sharing the struggling with herring DNA). A million thanks go to my supervisor Dorte for enlightening discussions and priceless advice on scientific (and baby) matters; it has been a really inspiring journey over the past six years, and I hope I am now ready to stand on my own feet. I also want to express my sincere gratitude to Jim and Lisa Seeb for taking me in (even though I was unwanted in the beginning), and introduce me to salmonids, Fred, Kerry, Lorenz, Sewall, Scott, Mette and a bunch of other cool and hospitable folks at

SAFS and around. Last, but not least, my fellow students Kristian, Mikkel, Diego, Sara, Henrik, the people you go to first when everything seems hopeless, or if you are just thirsty. The same accounts for Nina who has also been my extra sister and faithful partner the past years, whether we were swimming under the ice, road-tripping California or simply discussing smaller or larger things in life.

I have really enjoyed the times spent with you all, whether they concerned science, hair wax (Gary) or other important things in life.

I am grateful to my friends for being curious and understanding (or not) when I seemed more interested in working than partying for god knows which reason? My family (Far, Mor and Anna) for endless support and for thinking that I am a bright kid when I do not myself.

My two indisputable greatest motivation sources in life: Balder for being able to instantaneously cure a work related bad mood by reminding me about the true values of life, simply through your mere existence. Balder, just thinking about thinking about you makes me happy. Maria, I can't justify how much you mean for me in a few lines, but I am deeply grateful for your unconditional support via endless love and for giving me space and time, when my work really demanded it. Without you by my side, I could not have remained my sane self during these past years.

Silkeborg, December 2011

Morten Tønsberg Limborg

English abstract

Marine fishes represent a valuable resource for the global economy and food consumption. Accordingly, many species experience high levels of exploitation necessitating effective management plans. However, long term sustainability may be jeopardized from insufficient knowledge about intra-specific population structure and adaptive divergence. The large population sizes and high migration rates common to most marine fishes impede the differentiating effect of genetic drift, having led to expectations of no population structure and that the occurrence of local adaptation should be rare in these species.

Comprehensive genetic analyses on the small pelagic fish European sprat (*Sprattus sprattus*) revealed significant population structure throughout its distribution with an overall pattern of reduced connectivity across environmental transition zones. Population structure reflected both historical separations over glacial time scales and more recent colonisation of new habitats. Further, strong genetic divergence at several regional scales demonstrated limited connectivity among sea-going and local fjord populations along the Norwegian coast as well as indications for the potential of locally adapted populations in the brackish Baltic Sea.

If forces of natural selection are able to override the homogenizing effects of high gene flow, the detection of adaptive signatures has often been constrained by a general lack of genomic resources. However, advances in sequencing technologies now enable cost-effective developments of gene-associated markers facilitating detection of adaptive divergence. To further address the potential existence of locally adapted populations in small pelagic fishes, we developed hundreds of transcriptome derived markers to identify genes affected by natural selection in Atlantic herring (*Clupea harengus*). Comprehensive sampling throughout the northeastern Atlantic revealed clear genetic structure among regions, and coupled with environmental inference strong signatures of divergent selection at a range of candidate genes suggested adaptation to local temperature and salinity conditions.

A similar genome-scan based investigation of local adaptation was conducted in the salmonid *Oncorhynchus mykiss*. Despite profound socio-economic importance many populations have experienced strong declines and future conservation can be improved from identification of key environmental parameters and genes expected to maintain genetic diversity among populations. In contrast to marine fishes, salmonids are characterised by low gene flow, and together with the highly diverse habitats and phenotypes found among populations this suggest amble

potential for local adaptation to evolve. However, the genetic architecture and spatial scale of local adaptation is poorly known, and evidence has often been restricted to one or few genes at local scales. We found divergent selection for several genes often relating to local habitat conditions. Inference from known gene functions provided further evidence for adaptively important roles played by immune response genes.

Overall, results from this PhD revealed complex patterns of population structure and evidence for locally adapted populations in small pelagic fishes as well as interesting patterns of adaptively important candidate genes in a salmonid. These results contribute to our understanding of the evolutionary processes shaping biodiversity in the wild and findings may be extended from the actual species studied to assist managing fish resources under an evolutionarily sustainable framework in the future.

Dansk resumé

Marine fisk udgør en værdifuld økonomisk ressource på globalt plan, men mange arter udnyttes ikke på et bæredygtigt niveau, og bedre forvaltningsplaner er påkrævet. Utilstrækkelig viden om arters biologiske populationsstruktur og evolutionære tilpasning til lokale miljøer, kan medføre øget risiko for kollaps af lokale fiskebestande. Store populationsstørrelser og høje migrationsrater er typiske for marine fisk, hvilket i lang tid medførte en antagelse om at genetisk differentiering mellem populationer var usandsynlig i marine arter. Ydermere medfører disse faktorer en forventning om, at lokalt tilpassede populationer sjældent forekommer hos marine fisk.

Genetiske analyser af brisling (*Sprattus sprattus*) viste tydelig populationsstruktur over hele udbredelsesområdet med reduceret gen-flow over en række marine transitionszoner. Nogle populationer har sandsynligvis været isolerede fra før den sidste istid, mens andre har koloniseret deres nuværende områder på et senere tidspunkt. På regionalt plan viste analyser en skarp genetisk differentiering mellem populationer i Nordsøen og Østersøen i relation til en kraftig miljøgradient, hvilket kan tyde på, at populationer er lokalt tilpassede. Et andet studie viste, at populationer af brisling i norske fjorde kun i begrænset omfang blander sig med store havgående bestande, og at de således reagerer uafhængigt på fiskeri- og klimaeffekter.

Selvom naturlig selektion er stærk nok til at modvirke effekten af gen-flow, har det hidtil været svært at dokumentere graden af lokal tilpasning på grund af begrænsninger i tilgængelige genetiske metoder. Nye sekventeringsteknikker har gjort det muligt på én gang at udvikle mange gen-relaterede markører for nye arter, hvilket er et essentielt værktøj for at identificere lokalt tilpassede populationer. I et mere direkte forsøg på at detektere lokal tilpasning hos små pelagiske fisk, udviklede vi hundredvis af nye gen-relaterede markører for sild (*Clupea harengus*). Ved at analysere prøver fra hele Nordøst Atlanten og Østersøen fandt vi overordnet fire genetisk forskellige grupper af sild. Ydermere viste flere gener tegn på selektion i relation til lokale temperatur og salinitet forhold i overensstemmelse med lokalt tilpassede populationer.

Et lignende studie havde til hensigt at identificere kandidatgener for lokal tilpasning i regnbueørred (*Oncorhynchus mykiss*), som er en socioøkonomisk vigtig art i det meste af dens oprindelige udbredelse inklusiv det Nordvestlige Amerika. Klimaforandringer og menneskeskabte habitatforringelser har medført en stor tilbagegang af mange populationer, og i forbindelse med genopretningsplaner er øget viden omkring hvilke miljøparametre og gener, der

er afgørende for lokal tilpasning essentiel. I modsætning til marine fisk er regnbueørred karakteriseret ved meget store miljøforskelle og lavt gen-flow mellem populationer, hvilket sandsynligvis indebærer divergerende selektionsregimer mellem forskellige habitater med lokalt tilpassede populationer. Selvom flere resultater er i overensstemmelse med dette, ved man generelt meget lidt om det geografiske mønster eller den genetiske baggrund for sådanne tilpasninger. Vores resultater viste tydelige signaturer af divergerende selektion for flere gener, og ved at identificere de respektive geners biologiske funktion observerede vi en vigtig adaptiv rolle for forskellige typer af immunresponsgener.

De overordnede resultater fra denne PhD viste kompleks populationsstruktur og evidens for lokal tilpasning i små pelagiske sildefisk samt vigtige kandidatgener for tilpasning til lokale miljøer i regnbueørred. Disse resultater har øget vores viden om de evolutionære processer der opretholder biodiversitet i naturen, og de kan ligeledes medvirke til mere effektive forvaltningsplaner med henblik på evolutionært bæredygtige fiskerier i fremtiden.

Chapter 1

General introduction

Genetic population structure and local adaptation in fishes

Genetic variation reflects evolutionary processes

Genetic variation refers to polymorphic regions in the genome that can be scored for use in e.g. evolutionary studies of population and species histories. The origin of multiple alleles (polymorphism) at a given site in the genome originates from past **mutational** events. Once such polymorphisms exist in wild populations they are affected by a range of other evolutionary forces (Hedrick 2005a) where **gene-flow** reflects migration and leads to increased homogeneity among isolated populations. Contrary, **genetic drift** acting within populations leads to increased levels of differentiation among populations as a cause of random events between generations. Whereas the former two processes are considered neutral and expected to exert genome wide effects, imprints from **selection** are only expected to affect selected gene(s) and nearby linked genomic regions showing either increased (divergent selection) or reduced (balancing selection) levels of differentiation compared to neutrally evolving sites.

Neutral population structure

The time-scale upon which genomic imprints from the different processes accumulate among populations varies. For example, the time needed for new mutations to accumulate within single populations depends on the mutation rate and it may take several generations before such imprints become detectable between diverged populations. Contrary, signals from genetic drift, although depending on the effective population size (N_e), continuously accumulate over each generation between reproductively isolated populations. This leads to expectations of weak mutational imprints between recently diverged populations, and the relative mutational effect on genetic differentiation can thus be informative on the demographic history of a species (Excoffier et al. 1992; Pons and Petit 1996; *chapter 2*). The maintenance of contemporary neutral population structure mainly depends on the interplay between gene-flow and genetic drift acting over ecological time scales. Marine fish were traditionally perceived to represent genetically homogeneous populations since genetic drift is expected to exert little effect in species with large population sizes. This notion has now been largely abandoned from the accumulating evidence of biologically significant population structure in many marine fishes (Hauser and Carvalho 2008). Despite the often large population sizes, these results may reflect lower, than previously expected, levels of gene-flow among populations due to e.g. oceanographic retention reducing migratory potentials (Knutsen et al. 2011). Alternatively,

populations may potentially show adaptation to their local environments effectively reducing gene-flow from reduced reproductive fitness of immigrants.

Local adaptation

Evolutionary processes in the past have shaped contemporary genetic variation in extant species and populations in order to optimise their relative fitness within the environments to which they are exposed through natural selection. Similarly, environmental processes will continue to exert selective pressures at local populations in order to continuously optimise fitness in a changing habitat through evolutionary responses based on the standing genetic variation. The appearance of advantageous traits will thus develop over time through selective responses to changing environments. Whereas the nature of future responses of species to environmental change remains something between a black box and highly speculative predictions, current patterns of genetic variation within and among species provide us with a window towards understanding evolutionary processes in the past. In order to detect local adaptation, a well described neutral background of spatio-temporal population structure is of paramount importance to both evaluate the potential for local adaptation to occur (Hansen et al. 2002), and as a background upon which loci underlying selective pressures can be detected (Lewontin and Krakauer 1973; Storz 2005). However, lack of detectable neutral population structure does not necessarily preclude the existence of reproductive isolation between populations, which may be observed at genes under selection (Hemmer-Hansen et al. 2007a; *chapter 7*).

Why study population structure and local adaptation

Natural resources play a vital role in global economy alongside unprecedented rates of climatic and environmental changes with non-negligible contributions stemming from anthropogenic activities. Climatically driven temperature increases are expected to affect most ecosystems directly (e.g. species assemblages and physiological processes) and in-directly (e.g. through biotic interactions and changed hydrographical conditions). Recent studies have already shown wide ecosystem responses to anthropogenic disturbances coupled with climate changes (Bradshaw and Holzapfel 2006) including marine ecosystems (Perry et al. 2005; Casini et al. 2008; Dulvy et al. 2008). More specifically, increased temperatures are expected to alter salinity conditions in coastal and estuarine habitats from a combination of increased evaporation and freshwater runoffs leading to expected changes in species distributions and abundance (Mackenzie et al. 2007). Likewise, effects on freshwater systems are expected to include

changing river flow regimes and altered biotic interactions caused by species specific responses to environmental change (Wenger et al. 2011). Lastly, human activities including targeted harvest of specific traits such as size and growth rate (e.g. Jørgensen et al. 2007; Allendorf et al. 2008; Allendorf and Hard 2009) or habitat fragmentation like the building of dams (Waples et al. 2008) have substantial ecological and evolutionary impacts on wild animals.

In order to continuously improve conservation and management of biodiversity, studying evolutionary dynamics in varying and changing environments is of paramount importance for securing long term sustainability of wild animal resources (Schindler et al. 2010). Because, only through an increased understanding of past evolutionary processes can we aim at predicting future responses of different species in a changing world. These challenges laid the foundation for the European 7th Framework Programme project FishPopTrace (Box 1), in which I have been an active participant throughout my PhD period. In the following chapters I present my own contribution to the field from studies of two clupeids and a salmonid species.

Box 1. The FishPopTrace project

An increasing consumer demand coupled with limited natural resources have driven global fisheries into a state where more than two thirds of all exploited species are harvested outside safe biological limits (FAO 2010). Major problems for securing sustainable fisheries relate to Illegal, Unreported and Unregulated (IUU) fishing which have been estimated to represent at least $10 billion globally (Agnew et al. 2009), and which represent up to 25% of the global catch. Such unlawful actions include mislabelling of fish products in order to e.g. obtain a higher value or hiding origin of landings from endangered and protected populations (Miller and Mariani 2010). Contributing to these illegal activities is inadequate enforcement which lacks robust techniques for assigning fish and fish products to population of origin. Molecular markers capable of species-level identification and assigning individual fish to their biological population of origin (Ogden 2008) are particularly promising for counteracting IUU. However, a general challenge with marine fish is weak population diversification (Nielsen et al. 2009a). Despite an example relating to Atlantic cod (*Gadus morhua*) at large spatial scales (Nielsen et al. 2001), adequate population divergence and resulting statistical power for assigning individuals to population of origin at smaller spatial scales is commonly lacking for most marine fish.

The FishPopTrace project aimed to identify gene markers affected by diversifying selection in four commercially important species (Atlantic herring, common sole (*Solea solea*), European hake (*Merluccius merluccius*) and Atlantic cod) within EU managed waters (Martinsohn and Ogden 2008; 2009). These markers are intended to serve multiple purposes, including an increased understanding and resolution of the spatial structure of populations and how these may be adapted to local environments. Another goal was to use gene markers exhibiting high levels of population divergence to develop species specific panels with increased statistical assignment power.

To approach this goal a detailed strategy involving fish sampling, marker development, marker screening and validation of assignment power was incorporated in the FishPopTrace project (Martinsohn and Ogden 2009). First we developed transcriptome derived Single Nucleotide Polymorphism (SNP) markers in order to describe population structure and detect locally adapted populations. A subset of highly informative SNPs was then identified to develop a cost-effective tool box working throughout the food supply chain "from Ocean to fork" and intended to be used for assigning fish and fish products to population (and geographic area) of origin. Complementary strategies included microchemical and shape analyses of fish otoliths, fatty-acid profiles, proteomics and gene-expression profiles (see the FishPopTrace website: http://fishpoptrace.jrc.ec.europa.eu/ for more information) to further increase scientific evidence for detecting population of origin (Martinsohn and Ogden 2009). These results are expected to improve efficiency of fish forensic tools in order to better conserve marine resources through increased compliance towards fishery legislations and to minimise future levels of IUU through enforcement and deterrence.

Theory and methods

From genes to phenotypes and whole-organism performance

In the wild, natural selection favours individuals exhibiting the phenotypes best suited to the particular environment in which they live. Populations inhabiting varying environments may thus exhibit different phenotypes in order to optimise local fitness. When such phenotypes of adaptive importance have a heritable component (i.e. a genetic basis), natural selection will increase frequencies of the more adaptive alleles within populations and increase differentiation between populations. However, it is rarely possible to explain adaptive phenotypic variation from genetic polymorphisms alone since the gene only represents the first step in a cascade of complex biological levels ultimately shaping the phenotype of an organism (figure 1). Most population genetics studies of local adaptation in the wild are limited to data reflecting variation at the genetic level and often lack inference from several other levels of biological processes with potentially high impact on an organism's performance and relative fitness. Thus, in order to better link genetic variation to a certain measurable phenotype, a "mechanistic" approach considering multiple biological levels covering the entire pathway from genetic variation to whole-organism phenotype and fitness would be required (figure 1).

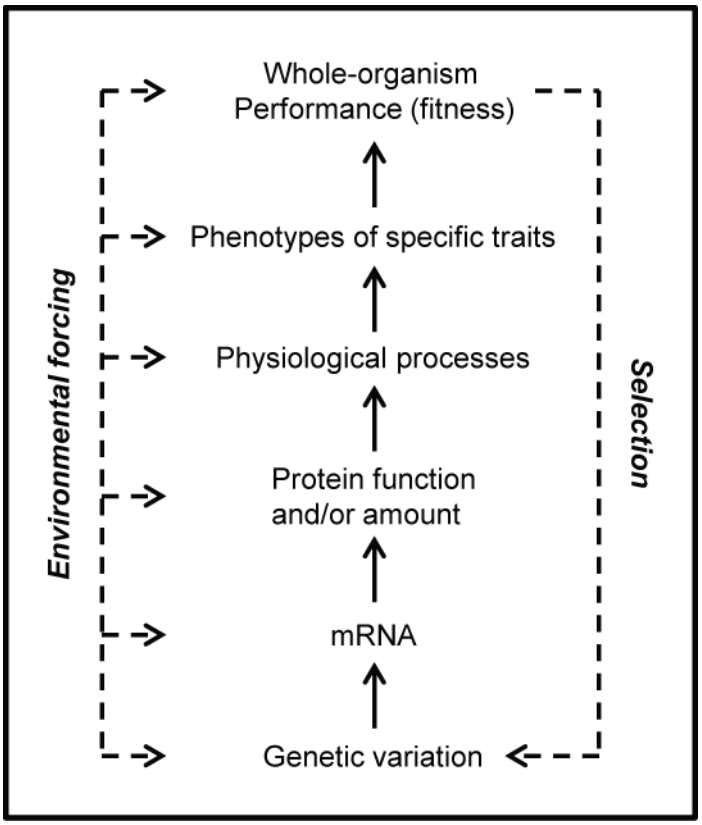


Figure 1 Diagram illustrating various levels of biological organisation that may, or may not, have a genetic basis affecting whole-organism fitness. Left broken line represents potential impacts on an organism from the external environment. Right broken line illustrates how natural selection may affect genetic variation in future generations by acting on the overall phenotype (figure modified from Dalziel et al. 2009).

While such approaches are only practical for few model[1] species (Dalziel et al. 2009) it can still be useful to apply a mechanistic perspective in studies of local adaptation in the wild (Rogers and Bernatchez 2007). Although only few steps in figure 1 can be addressed for nonmodel[2] species including most marine fishes, one can take advantage of the mechanistic perspective by considering existing knowledge about the effects of certain genes, proteins, physiological processes or other phenotypic traits suspected to play adaptive roles in the focal or related species. Such information may exist for some model species and can be useful in generating specific hypotheses to be tested at for example the genotypic level, or as *a posteriori* support for interpreting results indicating natural selection in the genome. One example of a well described model fish species which has had major impact for eco-evolutionary research in the marine realm is the threespine stickleback (*Gasterosteus aculeatus*) (e.g. Colosimo et al. 2005; Hohenlohe et al. 2010; Kitano et al. 2010). Thus, it has been recommended to continuously draw on well documented knowledge from key model species (ICES 2011 and references therein). It is further indisputable that the near future will provide massive resources enabling similar levels of detailed studies in organisms currently considered nonmodel species as for example exemplified by the recent publication of the Atlantic cod (*Gadus morhua*) genome (Star et al. 2011).

Despite the ever growing genomic resources available for nonmodel species, a particularly promising road to narrow the knowledge-gap in the genotype – phenotype pathway is the combined use of population genomics and quantitative genetics tools that aim at linking genetic variation with potentially adaptive, and quantifiable, phenotypic traits (Naish and Hard 2008; Stinchcombe and Hoekstra 2008). Despite the limited knowledge of especially the pathways including protein to physiological processes (figure 1) in most fish species, the combination of these tools operating at both the genetic and the phenotypic level appear especially promising for understanding genotype - phenotype interactions (Naish and Hard 2008). Generally, in order to increase the level of evidence for local adaptation in wild populations, it is an absolute necessity to consider multiple analytical approaches (see below) since no method alone can reveal information at all biological levels.

[1] Model organisms refer to extensively studied species often characterised by having extensive genomic resources like a fully sequenced genome

[2] Nonmodel organisms are here defined in a genetic framework of being restricted in terms of available molecular resources limiting genomic inference

Genetic markers for studying evolutionary processes

Molecular inference about population structure in marine fishes dawned in the 1960s where the application of allozyme markers demonstrated hitherto unknown population structure in a range of marine fish (Sick 1965b; 1965a; Nævdal 1968; Grant and Utter 1980). These studies effectively changed the perception of population structure in the wild by uncovering the ubiquity of intra-specific genetic variation in the wild. Many allozyme markers were assumingly affected by selection (e.g. Sick 1965a) challenging demographic inference, and subsequent technological improvements have made DNA based markers the preferred choice in most population genetic studies. Especially mitochondrial (mtDNA), microsatellite and Single Nucleotide Polymorphism (SNP) markers (see below) have revolutionized the field of evolutionary studies in the wild. Other types of markers include amplified fragment-length polymorphisms (AFLP), restriction fragment-length polymorphisms (RFLP) and randomly amplified polymorphic DNAs (RAPD), which have also had their heyday. However, due to drawbacks like dominance (i.e. homozygotes not being distinguishable from heterozygotes) and uncertainties relating to reproducibility of results, their usage has ceased (Schlötterer 2004), but see Avise (2004) for a more in depth discussion on the history of different types of molecular markers. In the work presented in this PhD I used the different attributes of mtDNA, microsatellites & SNP markers for answering questions relating to; population structure, demographic history, natural selection and candidate genes for local adaptation in two small pelagics and a salmonid fish species.

Historically, **mtDNA** markers have been the prime choice in most phylogeographic and population genetics studies owing to features such as high conservatism among species facilitating easy applications in new species using markers from close relatives (Avise et al. 1987). Mitochondrial DNA is haploid and generally maternally inherited, effectively reducing N_e of mtDNA markers to ¼ that of nuclear DNA. This means that genetic drift imposes a higher impact on mtDNA compared to nuclear DNA and that estimates of gene-flow only reflects female movements. Other general features of mtDNA include lack of recombination and selective neutrality, and even though many of these assumptions have been questioned in certain cases, mtDNA remains a valuable tool for studying phylogeography, particularly in species without large genomic resources (Galtier et al. 2009).

During the 1990s **microsatellite** markers became the single most popular genetic marker in ecological and evolutionary studies (Schlötterer 2004). This was to a large part due to attributes

like high mutation rates and concomitant high levels of polymorphism allowing unprecedented statistical power for detecting weak levels of population structure (Jarne and Lagoda 1996) as in many marine fish. The non-coding nature of most microsatellite motifs have led to a generally assumed neutrality of these markers (but see Nielsen et al. 2006) making them suitable for studying neutral evolutionary and demographic processes. However, the high mutation rates and polymorphic nature of microsatellites lead to potential drawbacks from size-homoplasy (Estoup et al. 2002) and bounding of the upper levels of population differentiation estimates such as F_{ST} (Hedrick 1999; 2005b). These issues have led to heated debate about the applicability of for example microsatellite derived F_{ST} estimates of population differentiation (see Holsinger and Weir 2009; Meirmans and Hedrick 2011 for two recent reviews and references therein). Microsatellites have, nevertheless, revolutionised the field of population genetics in marine, and other, fishes by uncovering hitherto undetected population structure in several species (Hauser and Carvalho 2008).

During the last decade we have seen a dramatic increase in the use of **SNPs** in studies of nonmodel organisms (Morin et al. 2004; Seeb et al. 2011a). SNPs refer to single base substitutions and thus represent the simplest and most abundant form of genetic variation in the genome. In combination with the recent development of next generation sequencing (NGS) techniques for large scale SNP discovery (Box 2), it is now possible to screen large numbers of SNPs in nonmodel species obtaining extensive genomic coverage compared to most previous microsatellite and mtDNA based studies (Luikart et al. 2003). Furthermore, by sequencing the transcriptome for SNP discovery, it has become more common to study markers from known genes and to reveal signatures of natural selection in nonmodel species (Bouck and Vision 2007). However, still in its infancy, a range of challenges including ascertainment bias and analytical considerations follow with the analyses of SNPs in nonmodel species, and these are reviewed in *chapter 5* of this thesis.

Despite each marker types' unique attributes for inferring phylogenetic relationships, weak population structure or signatures of selection, increased inference can be obtained from approaches combining the information that can be retrieved from different markers. In *chapter 2* I took advantage of such a multi marker approach by using mtDNA to infer phylogenetic relationships and comparing patterns of genetic differentiation with microsatellite markers, which allowed detection of weaker levels of structure not readily uncovered with mtDNA. Likewise, the use of markers under divergent selection (which are more readily discovered from large SNP

panels), may again enable identification of population differentiation undetectable with neutral markers (see *chapter 7*).

Box 2. From population genetics to genomics: The impact of Next Generation Sequencing

If population genetics refer to studies applying no more than ~20 markers for inferring population genetic patterns, population genomics can be defined as the genome-wide sampling of at least 100-1000s of markers to disentangle locus specific effects like selection from the genome wide influence of genetic drift and gene-flow (Luikart et al. 2003). Population genomics in nonmodel species has been made possible via NGS techniques (Shendure and Ji 2008), which have facilitated genomic scale studies in the wild (Ellegren 2008; Allendorf et al. 2010; Ekblom and Galindo 2011; Seeb et al. 2011a) including many fish species (Hauser and Seeb 2008). Genomic approaches now allow a range of hitherto unattainable inferences about the role of evolutionary forces in shaping genome wide variation. These include:

- Significantly improved accuracy in estimates of demographic parameters such as N_e and gene-flow from greatly reduced levels of inter-marker variation and the ability of filtering out non-neutral loci (Luikart et al. 2003).

- The possibility to effectively distinguish neutral from non-neutrally behaving markers in wild populations. Numerous studies have been carried out in human and model organisms (e.g. reviewed in Stinchcombe and Hoekstra 2008), but now applications to nonmodel organisms (e.g. Anderson et al. 2005; Eveno et al. 2008; Namroud et al. 2008) including fish (Moen et al. 2008; Bradbury et al. 2010) are increasing dramatically.

- Inferring the distribution and extent of neutral vs. selected genomic variation at the intra-genomic level (Nosil et al. 2009). For example, a recent study by Bradbury et al. (2010) found that 40 temperature related outlier loci out of more than 1600 screened SNPs only located to three out of 23 different linkage groups in Atlantic cod (*Gadus morhua*).

- An "environmental genomics" approach relating locus specific signatures of selection to an ecological context by considering associations with the surrounding environment (Landry and Aubin-Horth 2007).

- Increasing availability of new whole genome sequences of hitherto "nonmodel" fish species (e.g. Star et al. 2011) in the foreseeable future, allowing mapping 1000s of loci screened with high throughput genotyping techniques to known gene regions in the focal, or even closely related species (Sarropoulou et al. 2008). A fully sequenced genome may also allow targeted approaches screening a dense set of markers in a restricted part of the genome (reduced representation) for example containing candidate genes (Allendorf et al. 2010). Alternatively, picking a representative marker panel spanning the entire genome may significantly reduce the cost of screening for genome wide signatures of selection.

- Direct high throughput genotyping of a very large number of SNPs from randomly amplified DNA referred to as RAD-tag sequencing (Miller et al. 2007; Baird et al. 2008), as demonstrated in stickleback (Hohenlohe et al. 2010) and rainbow trout (*Oncorhynchus mykiss*) (Hohenlohe et al. 2011; Miller et al. 2011).

Methods for detecting local adaptation

Below, **divergent selection** refers to a scenario where individuals with a specific allele (A) experience higher fitness in a specific environment (X), while another allele (B) of the same gene is favoured in a different environment (Y). The frequency of the A allele will thus be increased through natural selection in populations adapted to the X environment in which the A allele has a selective advantage. Likewise, the B allele will be selected for in populations inhabiting the Y environment favouring individuals carrying the B allele (Kawecki and Ebert 2004). Compared to markers only shaped by neutral processes (see above), this leads to elevated levels of differentiation at this gene between populations adapted to X and Y environments, respectively, and this process is referred to as divergent (or disruptive) selection. Notably, alternative but similar scenarios include situations where only one of the alleles is under local selection while both alleles behave neutrally in remaining populations, however, this scenario leads to a similar pattern of increased differentiation at the respective gene (Kawecki and Ebert 2004). In the following I give a brief introduction to the most commonly used population genomics methods for detecting local adaptation through patterns of natural selection at specific genes. The objective here is to present the underlying principles and attributes of different methods for detecting local adaptation which I have considered during the work of my PhD. More exhaustive reviews of methods have been given elsewhere (e.g. Storz 2005; Vasemägi and Primmer 2005; Stinchcombe and Hoekstra 2008; Nielsen et al. 2009a).

Genome scans

The underlying principle of the genome scan approach rests on the hitch-hiking effect (Maynard Smith and Haigh 1974) implying that genetic markers linked to genes under selection will reflect the variation shaped by selective forces acting on a nearby functional gene. This is caused by linkage disequilibrium between genes under selection and the often effectively neutral flanking genomic regions, resulting from low recombination rates between the target of selection and flanking regions. Genetic markers located within such genomic regions affected by selection can then be statistically identified as initial signatures of selection. The first test developed for this purpose was the original method by Lewontin and Krakauer (1973), which compares single locus estimates of population differentiation (F_{ST}) among a set of population samples with a distribution of F_{ST} expected under neutral conditions. Loci subject to divergent selection are thus expected to show elevated levels of F_{ST} compared to the neutral distribution and similarly, loci under balancing selection (i.e. the same allele has a selective advantage in all populations) for

should show reduced levels of F_{ST}. Using the same principle, subsequent statistical refinements have been made to overcome previous short-comings of the method and a range of alternatives are now available (Beaumont and Nichols 1996; Vitalis et al. 2001; Beaumont and Balding 2004; Foll and Gaggiotti 2008; Excoffier et al. 2009).

Although this approach relies on linkage between markers and selected genes, this feature also limits the biological inference that can be drawn about detected outlier loci, since these are likely to reflect selection in nearby genes rather than being the actual target themselves. This is especially pertinent for anonymous markers like AFLPs and most microsatellites. One way to increase the chance of observing signatures of natural selection is to use markers (e.g. microsatellites or SNPs) in expressed regions of the genome (i.e. transcriptome) referred to as Expressed Sequence Tags (ESTs). These markers represent DNA that encodes information for subsequent protein synthesis (see figure 1) and as such are more likely to affect the ultimate phenotype potentially affected by selection (Bonin 2008; Namroud et al. 2008). Another drawback of the genome scan approach is that it requires a fairly large number of markers to obtain a robust neutral background upon which candidate markers for selection can be detected. Until recently, this limited the effective use of genome scans in most nonmodel organisms, but the advent of NGS (Box 2) has facilitated fast and cost-effective developments of large genomic resources (including EST markers) in nonmodel species (e.g. Barbazuk et al. 2007; Novaes et al. 2008; Hohenlohe et al. 2011; Milano et al. 2011). The increasing popularity of genome scans in population genomics is reflected in a recent boom of studies using this approach for both microsatellites (e.g. Martinez et al. 2011; Meier et al. 2011) and SNPs (e.g. Bonin et al. 2006; Moen et al. 2008; Namroud et al. 2008; Nielsen et al. 2009b; Bradbury et al. 2010). One crucial point to keep in mind is, that a detected “outlier” represents a marker falling outside a confidence interval (e.g. 95%) expected to encompass all neutral loci based on a statistical model. Underlying models differ among different genome scan methods, and markers obtaining outlier status in some methods may behave neutrally in others (Narum and Hess 2011). However, outliers detected by multiple methods based on different models effectively increase evidence of these markers being affected by selection (see e.g. *chapter 7*). Notwithstanding the inherent limitations of the genome scan approach given above and repeatedly pointed out (e.g. Kelley et al. 2006; Teshima et al. 2006; Excoffier et al. 2009; Hermisson 2009), the wide usage in recent years have highlighted different advantages and disadvantages for detecting signatures of selection in the wild. As a first step, genome scans can serve as excellent explorative tools for generating marker specific hypotheses and guide

downstream analyses (Beaumont 2005). By identifying and removing markers potentially affected by selection, genome scans can serve to compare neutral (by only including neutrally behaving markers) and adaptive processes in the wild (e.g. Gaggiotti et al. 2009; Nielsen et al. 2009c). Such assumingly neutral data sets are invaluable for estimating demographic parameters such as N_e and m, which are crucial indicators of the status of species or populations in conservation genomics (Luikart et al. 2003).

Candidate genes

Genes of known function expected to influence a phenotypic trait of adaptive importance can be considered as candidate genes for detecting local adaptation. Knowledge about the function of a gene can thus guide sampling of populations for example representing varying environments between which divergent selection would be expected. In combination with other inferences such as comparisons of phenotypes between environments, use of molecular tools can be applied to link variation in the gene with assumingly adaptive phenotypes or environmental conditions (Guinand et al. 2004). Such genes can then be subjected to a range of single-locus neutrality tests to infer if the gene deviates from neutral expectations in a fashion of divergent or balancing selection (reviewed in Ford 2002; Vasemägi and Primmer 2005). A range of these methods is furthermore demonstrated in an elegant candidate gene study of the Rhodopsin gene in relation to photic environments of the marine sand goby (*Pomatoschistus minutus*) (Larmuseau et al. 2009). Alternatively, the spatial distribution of genetic (allelic) variation at a candidate gene can be compared to that of a set of neutrally behaving markers. This was illustrated by Hemmer-Hansen et al. (2007) where the stress response gene *Hsc70* showed elevated levels of differentiation between populations of European flounder (*Platichthys flesus*) inhabiting different temperature and salinity regimes, compared to observed differentiation at assumingly neutral microsatellites. In *chapter 8* we used the Ewens-Watterson test (Ewens 1972; Watterson 1978) to look for balancing selection on a MHC class II gene in the salmonid *Oncorhynchus mykiss*[3], which is expected to maintain high levels of variation within populations due to its general role in immune response to external pathogens (Sommer 2005). Furthermore, the accelerated development of transcriptome derived and gene targeted markers in nonmodel species (e.g. Geraldes et al. 2011; Hemmer-Hansen et al. 2011; Seeb et al. 2011b), has led to an increased availability of markers residing within known genes, opposed to anonymous and non-coding markers. This increasing number of markers representing known genes will be a

[3] Throughout this chapter I refer to the Latin name of this species due to the existence of a myriad of popular names referring to specific life histories and evolutionary lineages (see below and *chapter 8*).

valuable resource for picking candidate genes in future studies drawing on findings from other species.

Landscape genetics

The field of landscape genetics aims at combining information from the surrounding "landscape" (e.g. geographic locations and environmental parameters like temperature, habitat type, etc.) with patterns of genetic variation among populations (Manel et al. 2003; Holderegger and Wagner 2006; Sork and Waits 2010). As the name implies, these methods have originally been developed for terrestrial systems but have already found great use in aquatic systems as well (Hansen and Hemmer-Hansen 2007; Selkoe et al. 2008). One group of methods investigates how environmental factors correlate with neutral genetic variation among populations, and if for example gene-flow appears to be limited across environmental barriers (Manni et al. 2004; Foll and Gaggiotti 2006; Faubet and Gaggiotti 2008). These methods have been applied to terrestrial (e.g. Heller et al. 2010) and a range of aquatic (Kenchington et al. 2006; Gaggiotti et al. 2009; Galarza et al. 2009; Gomez-Uchida et al. 2009) organisms for a limited number (<20) of microsatellite and mtDNA markers (see also *chapter 3*). The recent increase in number of available markers (such as SNPs), including markers representing adaptive genetic variation, is especially promising for coupling the surrounding landscape to functionally important genetic variation involved in local adaptation of wild populations (Holderegger et al. 2006). Methods often examine marker-environmental associations on a marker by marker basis and can potentially link environmental signatures of evolutionary importance to specific genomic regions (Joost et al. 2008; Coop et al. 2010). Several studies have already taken advantage of these new possibilities in nonmodel fish species (Nielsen et al. 2009b; Bradbury et al. 2010; Meier et al. 2011). These, and other studies, have contributed fundamental knowledge about, not only which factors are likely to drive local adaptation, but also about the spatial scales at which divergent selection operates in the wild. Notably, these methods are only correlative in nature, and a direct effect of tested factors on genetic variation often remains to be demonstrated un-equivocally. Nevertheless, results from these approaches still serve a useful hypothesis-generating role.

Other methods

In the course of my PhD I have mainly applied the three approaches described above, however, a range of other tools (e.g. QTL mapping, gene expression and proteomic based studies) to detect local adaptation have been applied in other nonmodel fish species and also deserve mention (see *chapter 9*). In common for these methods are that they all require complex laboratory facilities for breeding fish in common environments and/or large genomic coverage (preferably > 1000 markers). Many fish are not easily kept in captivity, and together with other constraints such as time and genomic resources, these approaches were not feasible during this PhD. However, consideration of these methods is extremely valuable for suggesting future directions of research based on the results obtained during my PhD, which is given in *chapter 9*.

Power of combining approaches for detecting local adaptation

The inherent uncertainty and limitation by any one method has repeatedly led to the advice of combining independent approaches (figure 2) for detecting local adaptation in order to effectively replicate results and strengthen conclusions (Luikart et al. 2003; Vasemägi and Primmer 2005; Stinchcombe and Hoekstra 2008; ICES 2011). Taking a hypothesis-testing approach by including gene markers expected to be under selection *a priori*, together with a large number of anonymous markers, increased evidence can be gained from combining the candidate gene approach with large scale genome scans. Thus, if the *a priori* candidates are suggested to be under selection in a genome scan combined with the known gene function qualifying it as a candidate, this effectively increases support for a model with selection.

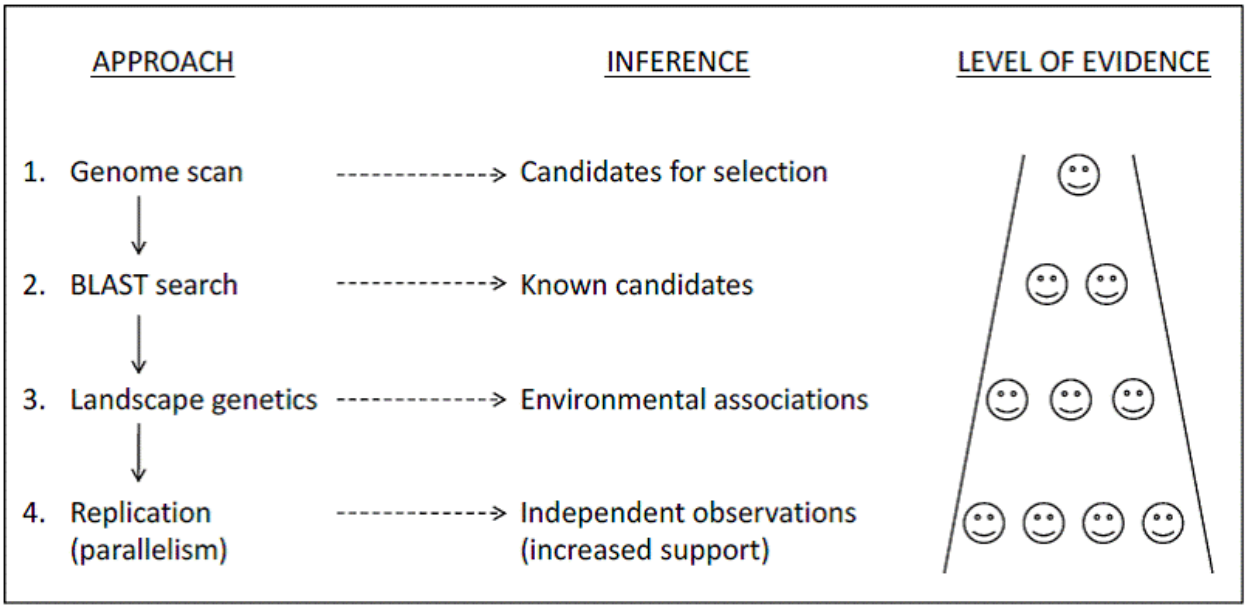


Figure 2 Conceptual diagram illustrating how combining inferences from multiple approaches can accumulate evidence for local adaptation in the wild (taken from ICES 2011). It should be noted that different approaches need not be performed in the order illustrated here, and that many other approaches for inferring local adaptation exist (see text).

Simultaneously, other markers, with no *a priori* expectations, may also show signatures of selection, and these genes may deserve more attention in downstream analyses. Such an approach was used by Nielsen et al. (2009b) who performed genome scans in Atlantic cod using 98 SNPs of which 18 were specifically designed to track candidate genes. Their results confirmed an adaptive status of some of these genes while also detecting a set of anonymous markers to be candidates for divergent selection. Co-workers and I followed a similar approach in a study on *O. mykiss* (*chapter 8*), where known gene identity for some outliers allowed us to conclude potential adaptive roles for MHC and interleukin immune response genes.

Another powerful combination of analytical approaches is to merge findings from genome scans with single marker based landscape genomic analyses. Genomic signatures of local adaptation are only expected to be found at genes or nearby genomic regions directly influencing fitness. Genetic markers within these regions will therefore be expected to show particularly strong correlations with environmental factors driving divergent selection between environments. This implies increased support for local adaptation when gene markers show signatures of selection from both genome scans and environmental correlations (see e.g. Nielsen et al. 2009b; Bradbury et al. 2010). Furthermore, by replicating analyses across independent environmental clines one may increase support for the evolutionary importance of a given environmental

parameter (Figure 2; Clarke 1975; Maggs et al. 2008). This was for example demonstrated by Hemmer-Hansen et al. (2007a) in populations of the euryhaline European flounder inhabiting geographically isolated low salinity environments. In *chapter 7* we also took advantage of this by comparing geographically independent clines of temperature and salinity in Atlantic herring (*Clupea harengus*), and in a similar way, we compared adaptive signals between different migratory life-history types in *O. mykiss* (*chapter 8*).

Spatio-temporal considerations for inferring adaptive variation

Adaptation to local environments is expected to reflect habitat heterogeneity, which may occur over relatively small geographic scales. This may compromise inference about the spatial scale at which selective forces are in play. This is particularly the case for widespread species such as most salmonids and marine fishes where spatially heterogeneous patterns of selection may be common (Miller et al. 2001; Hemmer-Hansen et al. 2007a). Thus, outliers detected from genome scans based on a broad geographic distribution of samples can be driven by just a few locally adapted populations and underlying selective processes may only occur at very local scales. To narrow down inference of selection in relation to local habitats, regional tests on a subset of local samples can be performed (Nielsen et al. 2009b; *chapters 7 and 8*). Alternatively, a hierarchical model approach can be taken in genome scan analyses (Excoffier et al. 2009) to infer whether outlier signals are driven solely by populations at local scales or if they represent more large scale patterns of selection among regional groups (see *chapter 8*).

Observed signatures of positive selection may either reflect on-going selection and/or imprints from past selective sweeps not yet eroded from populations through combined effects of recombination, gene-flow and drift (Garrigan and Hedrick 2003). Signatures from past selective sweeps are expected to remain longer in large N_e species (Garrigan and Hedrick 2003) like marine fish, implying great caution when interpreting contemporary signals of selection in the wild. Another phenomenon is the difference between local and global hitchhiking selection (Bierne 2010). Here local hitchhiking refers to a locally advantageous allele representing on-going divergent selection to local habitats, whereas global hitchhiking denotes a globally advantageous allele that is swept to fixation in the population of origin (i.e. past selection), and is in the process of spreading into neighbouring populations (Bierne 2010). To distinguish the latter two scenarios, a chromosome walk approach screening nearby variation can be applied, however this usually requires large genomic resources (Bierne 2010), which are currently unavailable for most marine fishes and other nonmodel species. Alternatively, observations of

similar adaptive signatures in geographically or phylogenetically distant populations may add further support for a model of on-going selection to local habitats (see discussion in Poulsen et al. 2011).

The study models

In the following I give a brief introduction to the studied groups of fish represented by the marine small pelagics European sprat (*Sprattus sprattus*) and Atlantic herring and the salmonid *O. mykiss*. It is not the aim here to comprehensively review the species' biology, which has been done elsewhere (e.g. Blaxter and Hunter 1982 for sprat and herring; Quinn 2005 for *O. mykiss*), but rather to introduce the specific biological characteristics making these species particularly interesting and suited for studying population structure and local adaptation in the wild.

Small pelagics (European sprat and Atlantic herring)

Small pelagics like sprat, herring (figure 3), sardines (*Sardinops sp.*), and anchovies (*Engraulis sp.*) are generally characterized by high cultural and economic importance (Whitehead 1985) and a crucial ecological role linking energy flow from lower trophic levels to top predators (Bakun 2006). This has also led to well described population structures in several small pelagics over the years (Carvalho and Hauser 1994; Hauser and Carvalho 2008), providing an invaluable knowledge base for guiding current and future research on local adaptation in the marine realm (Nielsen et al. 2009a). Small pelagics are representatives of what has been termed "classical marine fish", which is defined in an evolutionary context of having large N_e, broadcast spawning behaviour, large gene-flow potential due to high migration rates (m), and wide distributions (Palumbi 1994; Nielsen and Kenchington 2001). These attributes together with a general lack of physical migration barriers in the sea, have resulted in generally low levels of genetic differentiation among populations of marine fishes compared to anadromous (e.g. salmonids) and freshwater fishes (Ward et al. 1994; DeWoody and Avise 2000). While low levels of neutral population structure (reflecting demographic factors such as N_e and m) challenges detection of biologically significant differentiation among populations (Waples 1998), it facilitates detection of divergent selection. Notably, total response to selection (R) for a given trait in a population depends not only on the selection advantage (s), but also the size of N_e: $R=N_e*s$ (Robertson

1960). Thus, the combination of high N_e (low drift) and high gene-flow leading to low levels of neutral differentiation, and potentially strong adaptive responses to selection should make signals of divergent selection relatively easy to detect in marine fish (Foll and Gaggiotti 2008; Nielsen et al. 2009a). This, of course, assumes that gene-flow is not high enough to homogenise allele frequencies among populations between every generation. Herring is known to exert strong homing behaviour to natal spawning grounds (Iles and Sinclair 1982; Aro 1989) despite extensive feeding migrations and mixing with other populations (Ruzzante et al. 2006), which indeed suggest limited gene-flow and potential for local adaptation.

Figure 3 School of herring (photo: *unknown*).

Another feature of fishes in general is their exothermic biology leading to profound physiological responses to changes in external factors such as temperature, salinity or oxygen content. When species or populations are not capable of responding to environmental changes through either phenotypic plasticity or spatial movements, they may either go extinct or adapt through evolutionary response over few generations. The wide distributions of most classical marine fish often encompass highly heterogeneous environments with spatially varying selection pressures elegantly setting the scene for local adaptation to evolve (Palumbi 1994; Kawecki and Ebert 2004) and locally advantageous alleles may quickly sweep to high frequencies within populations. In sprat, three different sub-species have been described (Whitehead 1985) suggesting old divergence and low mixing among contemporary populations, which may have adapted to their local habitats. All these features make small pelagics excellent models for

studying population structure and local adaptation with a focus on environment – genome interactions. This line of research has been termed “environmental genomics” (Cossins and Crawford 2005), and represent a promising field for increasing our evolutionary sense of local adaptation in the sea.

Salmonids (*Oncorhynchus mykiss*)

Salmonid fishes have played a key role in evolutionary studies for several decades due to their extensive geographic distributions, immense life-history diversity and strong natal homing behaviour to mention a few attributes (Hendry and Stearns 2004; Quinn 2005). Salmonid species of the genus *Oncorhynchus* are naturally distributed in the North Pacific Ocean and spawn in freshwater tributaries from northeast Asia over Alaska down to Mexico on the west coast of North America (MacCrimmon 1971; Utter et al. 1980). The species *O. mykiss* (figure 4) has attracted particular attention as a function of its value in recreational and aquaculture activities coupled to its enigmatic biology (Halverson 2010). Especially its importance to aquaculture has led to the development of extensive genomic resources (e.g. Thorgaard et al. 2002; Miller et al. 2011), which also comprise an excellent tool box for studying local adaptation in the wild (e.g. Narum et al. 2010; Miller et al. 2011).

Figure 4 *Oncorhynchus mykiss.* (Photo kindly provided with permission from Daniela & Benno Wolf (©www.taucher.li).

Two life-history forms exist; the anadromous *steelhead* which performs extensive ocean-going migration before returning to natal spawning areas, and the resident *rainbow trout* which spends its entire life cycle in freshwater. The evolutionary significance of these alternative life-styles has

been extensively studied (Hendry et al. 2003), but evidence for a genetic component identified through candidate genes associated with life-history remains sparse (but see examples in Narum et al. 2011 and *chapter 8*).

After the last glacial maxima previously isolated lineages of *O. mykiss* came into secondary contact during the re-colonisation of the Pacific Northwest (*chapter 8*). This included populations within lineages colonising different environments and populations between lineages colonising similar environments (e.g. Miller et al. 2011). This aspect makes *O. mykiss* an ideal model for distinguishing neutral from adaptive evolutionary processes since natural replicates of phylogenetically distinct populations performing both migratory life-styles or inhabiting similar environments are abundant throughout its native distribution.

Major objectives and discussion of results

The overarching goal of this PhD was to describe population structure and detect local adaptation in marine small pelagics, and to shed light on the evolutionary factors more likely to explain adaptive signatures between populations. For comparison, insights from a different model system, the salmonid *O. mykiss*, are given in *chapter 8*, presenting a similar approach for detecting signatures of local adaption as applied for herring in *chapter 7*. The methods applied throughout the different studies reflect the ongoing transition from a population genetics to a population genomic era within the field of molecular ecology (Ellegren 2008; Nielsen et al. 2009a; Allendorf et al. 2010; Ouborg et al. 2010). In the following I present the main research questions, strategies and results of the work conducted during my PhD. Only main findings are discussed here in order to tie major objectives and accomplishments, while more in depth discussions of results are given within each of the manuscripts presented in the following *chapters 2-8*. Concrete steps to follow up on findings from this PhD are also given here, whereas a wider perspective on future directions for studies on local adaptation in fish and other nonmodel species is given in *chapter 9*.

Understanding demographic history and neutral population structure in the sea

Knowledge about species' demographic history and neutral population structure is essential for assessing the "potential" for local adaptation to occur within species (Hansen et al. 2002), and

for designing studies aiming at detecting locally adapted populations in the wild. For example, local adaptation is less likely to occur between populations where high gene-flow completely homogenise genetic variation impeding the effects of divergent selection. Contrary, local adaptation is more likely to evolve in large N_e populations where selective forces are expected to more effectively overrule the confounding effects of genetic drift. Neutral population structure is relatively well described in herring (Bekkevold et al. 2005; Jørgensen et al. 2005; Mariani et al. 2005) and *O. mykiss* (e.g. Allendorf and Utter 1979; Utter et al. 1980; McCusker et al. 2000), but not in sprat. *Chapters 2*, *3* and *4* report studies of population structure in sprat. Little molecular evidence about population structure within this species existed (e.g. Nævdal 1968), until recently where a comprehensive phylogeographic study was presented by Debes et al. (2008). Using mtDNA they showed the existence of two major evolutionary clades contradicting with previous categorization of sub-species (Whitehead 1985), demonstrating the complementary power of molecular approaches to morphometric characters in elucidating population structure. The work by Debes et al. (2008) was conducted in parallel with my microsatellite based study on population structure presented in *chapter 3*, and in *chapter 2* we collaboratively combined mtDNA and microsatellite data to investigate the demographic history of sprat. In this thesis the order of *chapters 2*, *3* and *4*, presenting population genetic analyses in sprat reflect the evolutionary time scale inferred from the respective results rather than a chronological order of publication.

In *chapter 2* we found that relative effects of mutation and genetic drift in explaining differentiation between populations varied among four marine transition zones with a tendency towards a weaker mutational impact at higher latitudes. This suggests more recent divergence of populations within the most northerly distribution in accordance with a general pattern across organisms of more recent colonisation after the last glacial maximum (Hewitt 2000). Overall, we detected four genetic clusters corresponding with the occurrence of independent population units in areas separated by transition zones. These results cannot rule out the existence of finer scale population structure within these areas. However, in *chapter 4* we addressed this question by performing denser regional sampling in the northeast Atlantic.

In *chapter 3* we investigated the population structure of sprat in its northern distribution and our results revealed overall patterns of genetic homogeneity within both the North Sea and Baltic Sea regions, but a steep genetic division between the two seas, describing a strong environmental cline characterised by drastic changes in salinity. Similar overall patterns of strong genetic differentiation across this environmental cline has been reported in a range of

other marine fishes (Nielsen et al. 2003; Nielsen et al. 2004; Bekkevold et al. 2005; Hemmer-Hansen et al. 2007b) suggesting the occurrence of a multi-species hybrid zone. However, a comparison of genetic structures among species revealed interesting differences in terms of the spatial location showing the steepest genetic divides (*chapter 3*). Such inter-specific comparisons appear promising for identifying key traits determining when and where population boundaries may evolve by considering ecological characteristics of different species in relation to genetic patterns.

The results presented in *chapter 4* aimed at determining population structure in sprat at its northernmost distribution with emphasis on local fjord populations along the Norwegian coast. The results revealed a general pattern of limited connectivity between sea-going sprat and a more genetically homogeneous group represented by all sampled fjord populations. These findings are in line with an observed discordance between sprat abundance in fjords and the North Sea region (see discussion in *chapter 4*) suggesting the existence of at least two different groups acting as independent demes under an ecological paradigm (*sensu* Waples and Gaggiotti 2006).

Altogether, we found highly significant neutral population structure throughout the distribution of sprat at a species (*chapter 2*) and regional (*chapters 3* and *4*) scales, that may reflect locally adapted populations. The extent and nature of such adaptations should be targeted in future efforts which could furthermore benefit by including comparisons with other species for increasing knowledge on key traits affected by natural selection in the sea.

Applying SNPs in nonmodel organisms; opportunities and challenges

The field of molecular ecology in nonmodel organisms, including most fish, has mainly relied on population genetics approaches often applying 10-20 highly diverse microsatellite markers (Luikart et al. 2003). It was therefore necessary to thoroughly consider analytical issues before taking the next step towards population genomics applying more, but less diverse, markers like SNPs. The often bi-allelic nature of these types of markers reflects substantial differences in their mutational pattern and levels of variation compared to e.g. microsatellite markers. Secondly, new population genomic data sets often consist of marker numbers in the hundreds or thousands, thus limitations of classical softwares for conducting statistical tests may challenge analyses of genomic data sets. Even though statistical improvements are expected to

quickly follow the new demands, it seemed timely to evaluate the proposed wonders (Box 2) and potential drawbacks for applying large scale SNP data for studies on nonmodel species in general.

In *chapter 5* we therefore reviewed the most relevant population genetic softwares for suitability to handle large scale SNP data sets. Compared to other recent reviews about SNPs dealing with either the general applicability in ecology and evolution (e.g. Morin et al. 2004) or more technical aspects (e.g. Garvin et al. 2010), our review took an analytical perspective addressing a range of challenges relating to data analyses.

From this review we concluded that a solid tool box of statistical tools for analysing large SNP data sets already exists, with many more being developed at a pace implying that the availability of statistical tools is not likely to become a limited factor in current population genomic studies of nonmodel organisms. However, alongside the continued development of sequencing and genotyping techniques, dealing with analytical challenges relating to genomic data sets are expected to remain an everyday job for a time to come. That being said, the importance of conducting new empirical studies applying genomic tools currently available are urgently needed for identification of the accompanying challenges. A continuous usage is thus crucial for securing timely solutions addressing such challenges while preparing molecular ecologists for even larger data sets in the future.

A new population genomic resource for Atlantic herring

In order to distinguish gene regions conforming to neutral or non-neutral processes of evolution, a large marker panel is warranted for obtaining high genomic coverage to detect signatures of selection. Using expressed DNA regions for marker development, one significantly increases the chance of finding signatures of selection at functional genetic variation affecting fitness of individual phenotypes. In order to accomplish this we used NGS technology to sequence the transcriptome of eight herring individuals sampled throughout the distribution to detect and develop working SNP assays (*chapter 6*).

Although this strategy has been demonstrated in other organisms (Barbazuk et al. 2007; Hyten et al. 2010) it is not without challenges when applied to a new species with little or no existing genomic resources like herring. The lack of a reference genome implies that a *de novo*

assembly of single sequence reads into longer consensus contig groups was prepared to act as sequence backbones for downstream SNP discovery. This led to the choice of using 454 GS FLX sequencing (Roche) due to the advantage from the longer read lengths provided by this platform significantly improving assembly (Metzker 2010). Recalling that one of our objectives in the FishPopTrace project was to develop informative SNPs for both discriminating among populations and detecting local adaptation (Box 1), we applied a transcriptome based approach thereby focusing on SNPs in gene regions, which are more likely to be affected by selection (*chapter 6*). However, this approach often comes with the cost of an increased false positive rate of *in silico* detected SNPs due to for example the inclusion of undetected intron-exon boundaries within flanking regions in some fraction of SNPs (Lepoittevin et al. 2010). Alternatively, sequencing genomic DNA significantly reduces this risk (see e.g. Hyten et al. 2010), but here the majority of SNPs is expected to reside within anonymous and non-transcribed regions most likely representing neutral genetic variation. This may, however, be preferable for developing SNP panels designed for estimation of neutral demographic parameters.

In *chapter 6* we used a NGS approach to successfully identify hundreds of SNPs for use in downstream genomic analyses of herring. However, similar future efforts will have to consider a range of choices relating to for example; i) type(s) of sequencing platform used, ii) sequence cDNA or genomic DNA libraries, and iii) effort put into the validation pipeline, all of which depend on the concrete research questions being asked and available resources. Despite inherent pros and cons of different NGS approaches, future studies will inevitably benefit from the continued technological improvements offering both higher numbers and increased lengths of reads improving the quality of *de novo* assemblies (Ekblom and Galindo 2011).

Genomic signatures of local adaptation

Herring

It is now well known that most classical marine fish to some degree exhibit biologically significant population structure as inferred from neutral genetic markers (Hauser and Carvalho 2008). Temporal stability of such patterns may be upheld by either physical barriers hindering gene-flow between demes, increased fitness of locally adapted populations minimising the reproductive success of immigrants (further reducing gene-flow), or a combination of both.

However, the relative effect of these neutral and selective scenarios has hitherto been difficult to assess in classical marine fish due to lack of genomic coverage (but see Larsen et al. 2007; Bradbury et al. 2010).

In *chapter 7* I used our new genomic resource for herring (*chapter 6*) to take the next step and distinguish spatial patterns of neutral and selected variation among gene associated markers. To my knowledge, this represents the hitherto most comprehensive genome scan in a nonmodel classical marine fish, disregarding a recent study by Bradbury et al. (2010) on Atlantic cod which can be argued to represent a "semi-model" organism after the recent publication of its genome sequence (Star et al. 2011). Furthermore, coupled with strong signatures of environmental effects (especially characterised by low salinity) driving local adaptation in some populations, we identified a hitherto unprecedented number of candidate genes in this species. Known gene functions in some non-synonymous candidate SNPs included haemoglobin and heat shock proteins and such information reinforce the evidence for adaptive functions of these genes (*chapter 7*).

Genomic resources for Atlantic herring and related species will inevitably increase in the future, and these will facilitate mapping and annotation of most candidate genes allowing a fuller understanding of the genetic architecture underlying adaptively important traits. Such approaches have for example already identified few multi-gene regions of assumingly important adaptive roles in other marine fishes (Bradbury et al. 2010; Hohenlohe et al. 2010).

We found significant and spatially replicated correlations for a range of candidate genes in relation to temperature and salinity (*chapter 7*). In order to draw a more direct link between the environment and genes affecting overall fitness, our results can be seen as hypothesis generating for testing the effects of temperature and salinity on divergently adapted populations kept under controlled conditions. Such common garden experiments may be challenging due to the difficulty of rearing many marine fish in captivity. However, experiments performed on wild caught individuals have proven rewarding in a number of marine species (Nissling and Westin 1997; Larsen et al. 2007) including Pacific herring (*Clupea pallasi*) (Griffin et al. 1998) and sprat (Petereit et al. 2008). Future efforts combining inference from candidate genes with investigations of fitness effects in controlled environmental settings are indeed appealing for coupling genetic signatures with phenotypic responses.

Herring constitute a potentially interesting model for studying a metapopulation (*sensu* Hanski 1998) scenario in the Baltic – North Sea transition zone and western Baltic Sea aiming at identifying potential source-sink population dynamics over time. For example, Bekkevold et al. (2007) demonstrated different modes of origin in two small regional herring contingents exhibiting shifted spawning times compared to larger sympatric components. This suggest a complex pattern of multiple minor herring components acting more or less independently from larger assumed source populations, at least on an ecological time scale (Waples and Gaggiotti 2006). Future genomic approaches could take advantage of increased power for distinguishing local components of often weakly differentiated herring groups by including genetic markers under temporally stable divergent selection (Nielsen et al. *unpublished manuscript*).

Lastly, improved resolution of herring population structure over time and space from both neutral and adaptive markers may serve as an important component in a multidisciplinary approach combining genetic structure with demographic data on abundance and distribution in order to elucidate a potential meta-population stabilising effect of population diversity (see Schindler et al. 2010 for a similar scenario in a salmonid system).

Oncorhynchus mykiss

In *chapter 8*, I applied a similar genome scan approach to investigate large-scale patterns of natural selection and identify candidate genes in *O. mykiss*. While the detection of outliers for divergent selection in classical marine fish is facilitated by a weak background of neutral differentiation (Nielsen et al. 2009a), salmonids are generally characterised by higher levels of neutral structure caused by smaller N_e and lower levels of gene-flow (DeWoody and Avise 2000). These differences between marine and salmonid fishes may be translated into expectations of increased type I errors (i.e. more false positive outliers) in marine fish whereas an opposite pattern of high type II errors (i.e. more false negatives) is likely to prevail in salmonid genome scans. In both cases, this warrants great care in the interpretation of the biological significance underlying statistically significant outliers from genome scans (Fraser et al. 2011). Nevertheless, we found several interesting signatures of evolutionary important genes and traits in *O. mykiss* (*chapter 8*).

By combining inference from genome scans with landscape models we found strong evidence for divergent selection at varying spatial scales including different habitats and phylogeographic lineages. Together with a recent study by Narum et al. (2010), our landscape analyses contribute to recent evidence for temperature driven selection between habitats in *O. mykiss*,

which may be related to temperature-induced differences in pathogenic communities (Tonteri et al. 2010) and/or accelerated growth rates in populations adapted to cold temperatures (Miller et al. 2011).

We identified candidate genes for being under divergent selection between populations exhibiting resident or anadromous life-styles in *O. mykiss*, which is in accordance with two recent studies specifically focusing on selective signatures relating to migratory behaviour (Martinez et al. 2011; Narum et al. 2011). However, pure genome scan derived findings are only weak evidence, and further investigations should for example focus on more direct links between physiological processes related to smoltification (and anadromy) and potential candidate genes. Such investigations could benefit from a combined approach also considering inference from the transcriptome and proteome levels (see e.g. Giger et al. 2006; *chapter 9*).

Furthermore, merging evidence from independent genome scans, single gene neutrality tests, and earlier findings suggested balancing selection at a class II Major Histocompatibility Complex (MHC) gene and divergent selection at multiple interleukin genes. To gain further insight about the selective processes maintaining high levels of functional variation at MHC genes, a more direct coupling between genetic variants and for example parasite and pathogenic load should be made, as for example demonstrated by McCairns et al. (2011) in stickleback. It will also be interesting to further assess the role of interleukin gene responses in wild populations adapted to different pathogenic habitats in order to shed more light on the potential adaptive role played by these genes in wild *O. mykiss* populations. This could for example be addressed with a common garden set up monitoring population specific gene expression at interleukin genes, which are already being used as indicators of innate immune responses towards various pathogens in *O. mykiss* (Klaper et al. 2010; MacKenzie et al. 2010). The MHC represents a well-studied gene-family in fishes and other vertebrates with a long line of evidence for adaptation to local environments (Bernatchez and Landry 2003; Sommer 2005). Contrary, the adaptive role of interleukin genes, which are involved in early responses of the innate immune system (Secombes et al. 2011) have not received similar levels of attention (but see e.g. Narum et al. 2011). Our results contribute to the accumulating evidence that immune-relevant loci in general play an important adaptive role in wild salmonid populations (e.g. Tonteri et al. 2010).

Closing remark

In conclusion, by combining inferences from a range of different methodological approaches we found convincing signatures of natural selection related to local environments among wild populations of herring and *O. mykiss*. However, much remains to be done in order to fully comprehend the selective processes shaping adaptive divergence in space and time, but the results presented in *chapters 7* and *8* contribute to this understanding and will help guide future efforts within the field.

Background and objectives of manuscripts

In the following I give a brief description of the projects and motives leading to the different studies presented in the following chapters. It is not the intention here to present and discuss results, which have been done elsewhere. Rather, I present the objectives for each study in an evolutionary framework supposed to link the relevance of each study to an overarching goal of understanding the underlying mechanisms of population structure and local adaptation by learning from marine and salmonid fishes. Supplementary files can be found on web sites of the respective journals.

Chapter 2 Imprints from genetic drift and mutation imply relative divergence times across marine transition zones in a Pan European small pelagic fish (Sprattus sprattus)

This study is the result of a collaborative effort in a small group of population geneticists representing a working group on genetic biodiversity (GBIRM[4]) of marine organisms in European waters under the EU network of excellence MARBEF[5]. The major objective of this study was to uncover the demographic history of sprat by using known transition zone areas as reference points for past population splitting events. A distribution wide data set of sprat was produced to compare genetic imprints at two different marker systems (mtDNA and microsatellites). The wide geographic sampling of sprat enabled interesting comparisons from signatures of population differentiation across several known multi-species transition zones reflecting diverse historical events. This study is published in *Heredity* and included here with permission from the publisher Nature Publishing Group.

Chapter 3 Genetic population structure of European sprat (Sprattus sprattus L.): differentiation across a steep environmental gradient in a small pelagic fish

This study represents a major part of the work carried out in connection with my master thesis (Limborg 2007) although substantial re-writing of the manuscript and the entire publication process were undertaken during my PhD. As such, the article does not represent work for my PhD thesis but is nonetheless included here because of its relevance for linking the remaining work carried out during my PhD and presented in this thesis. The results add insight to the clear genetic structuring observed for multiple marine fishes across the strong environmental cline connecting the North Sea and Baltic Sea, demonstrating biologically significant intra-specific

[4] http://www.marbef.org/projects/gbirm/index.php
[5] http://www.marbef.org

biodiversity despite the high N_e and high gene-flow in small pelagic species. This study is published in *Marine Ecology-Progress Series* and included here with permission from the publisher Inter-Research.

Chapter 4 Microsatellite DNA reveals population genetic differentiation among sprat (Sprattus sprattus) sampled throughout the Northeast Atlantic, including Norwegian fjords

Local Norwegian fjord populations of sprat have shown large fluctuations in abundance with extended periods of low levels, despite seemingly large and stable oceanic populations in the North Sea region. This fact led to the current study initiated by scientists from the Institute of Marine Research (IMR) in Norway to which I contributed with sampling, microsatellite data, statistical analysis and interpretation of results. The article investigates population genetic structure in sprat from the northernmost parts of its distribution including samples from several isolated Norwegian fjords and the adjacent North Sea. This study is published in *ICES Journal of Marine Science* and included here with permission from the publisher Oxford Journals.

Chapter 5 Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges

SNPs gain increasing popularity for studies in molecular ecology and for detecting signatures of natural selection in particular. However, their usage in nonmodel species comes with an initial set of challenges like the high number of markers often applied, ascertainment bias, and inclusion of non-neutral loci which need careful consideration before using them in population genetic studies. This review considers these issues and represents the result of a collaborative effort of the FishPopTrace consortium, where we thoroughly discussed such issues at a work-shop. This review is published in *Molecular Ecology Resources* and included here with permission from the publisher John Wiley and Sons.

Chapter 6 SNP discovery using next generation transcriptomic sequencing in Atlantic Herring (Clupea harengus)

This study aimed at developing a large panel of SNP assays for herring as no major genomic resources were available for this species at the launch of the FishPopTrace project. We sampled herring from major parts of its Northeast Atlantic distribution for high throughput sequencing in order to discover new SNPs for downstream studies. This work included a range of challenges from study design over analyses of sequence data to final validation of SNP

assays via high throughput genotyping. Therefore, many people were included in order to obtain adequate expertise for the different steps in the study. The final accomplishment represents one of the large assets of the FishPopTrace project bringing together a diverse set of experts synergistically generating unique output unattainable by any single research group within the project. This study was conducted jointly and in parallel with the now published SNP resource for European hake (*Merluccius merluccius*) by Milano et al. (2011), explaining the similarity in approach between these two studies. This study is published in *PLoS ONE*.

Chapter 7 Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (Clupea harengus)

One of the main objectives of FishPopTrace was to increase our understanding about the underlying evolutionary processes shaping neutral as well as adaptive genetic variation within species. The occurrence of locally adapted herring populations has been suggested based on previous studies describing highly significant patterns of population structure in this species. This will also be in accordance with its successful colonisation of extreme marine ecosystems like the northernmost Baltic Sea characterised by near freshwater, unlike most other marine fishes (EEA 2002). In this study we used the SNP resource developed in *chapter 6* to shed light on the adaptive side of population structure in herring. We aimed at distinguishing distinct patterns of neutral from selected variation in population structuring. Further, we used a landscape genomics approach to identify both candidate genes for being under selection and key environmental variables exerting local selective pressures. This study is published in *Molecular Ecology* and included here with permission from the publisher John Wiley and Sons. Please note that an Erratum is now accompanying the online version of this publication.

Chapter 8 Signatures of natural selection among lineages and habitats in Oncorhynchus mykiss

As part of the DTU PhD programme, students are encouraged to spend part of their time with an external research group within the field in order to experience different research environments while gaining experiences complimenting the expertise of their own institution. When deciding where to go I prioritised groups with a longer history of, and experience with, the application of SNPs for studying evolution in fishes. To fulfil these requirements I went to study at the School of Aquatic and Fishery Sciences at the University of Washington, Seattle, USA June-October 2010, where I worked with Fred Utter and the group of Jim and Lisa Seeb[6]. Here,

[6] http://fish.washington.edu/research/ipseg

I continued my research on local adaptation in fishes adding a salmonid twist. I used a newly developed SNP panel for investigating signatures of selection in *O. mykiss* representing a well described and complex salmonid species system with distinct evolutionary lineages and life-history strategies. Genetic evidence for natural selection at the genome level is still scarce for *O. mykiss* in the wild, and my study represents one of the first investigations of large scale selective signatures applying hundreds of markers in a genome scan approach. This study is published in *Ecology and Evolution*.

References

Agnew, D. J., J. Pearce, G. Pramod, T. Peatman, R. Watson, J. R. Beddington and T. J. Pitcher (2009). Estimating the worldwide extent of illegal fishing. *PLoS ONE* **4**: e4570.

Allendorf, F. W., P. R. England, G. Luikart, P. A. Ritchie and N. Ryman (2008). Genetic effects of harvest on wild animal populations. *Trends in Ecology & Evolution* **23**: 327-337.

Allendorf, F. W. and J. J. Hard (2009). Human-induced evolution caused by unnatural selection through harvest of wild animals. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 9987-9994.

Allendorf, F. W., P. A. Hohenlohe and G. Luikart (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11**: 697-709.

Allendorf, F. W. and F. M. Utter (1979). Population genetics.Pp. 407-454 *in* W. S. Hoar, D. J. Randall and J. R. Brett, eds. Fish Physiology Volume VIII. Academic Press, New York **VIII**.

Anderson, T. J. C., S. Nair, D. Sudimack, J. T. Williams, M. Mayxay, P. N. Newton, J. P. Guthmann, F. M. Smithuis, T. T. Hien, I. V. F. van den Broek, et al. (2005). Geographical distribution of selected and putatively neutral SNPs in Southeast Asian malaria parasites. *Molecular Biology and Evolution* **22**: 2362-2374.

Aro, E. (1989). A review of fish migration patterns in the Baltic. *Rapports et Procés-Verbaux Des Réunions Du Conseil International Pour l'Exploration de la Mer* **190**: 72-96.

Avise, J. C. (2004). Molecular markers, natural history and evolution. Sunderland, MA, USA, Sinauer Associates, Inc.

Avise, J. C., J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb and N. C. Saunders (1987). Intraspecific phylogeography - the mitochondrial-DNA bridge between population-genetics and systematics. *Annual Review of Ecology and Systematics* **18**: 489-522.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko and E. A. Johnson (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**: e3376.

Bakun, A. (2006). Wasp-waist populations and marine ecosystem dynamics: Navigating the "predator pit" topographies. *Progress in Oceanography* **68**: 271-288.

Barbazuk, W. B., S. J. Emrich, H. D. Chen, L. Li and P. S. Schnable (2007). SNP discovery via 454 transcriptome sequencing. *Plant Journal* **51**: 910-918.

Beaumont, M. A. (2005). Adaptation and speciation: what can F_{ST} tell us? *Trends in Ecology & Evolution* **20**: 435-440.

Beaumont, M. A. and D. J. Balding (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**: 969-980.

Beaumont, M. A. and R. A. Nichols (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B-Biological Sciences* **263**: 1619-1626.

Bekkevold, D., C. Andre, T. G. Dahlgren, L. A. W. Clausen, E. Torstensen, H. Mosegaard, G. R. Carvalho, T. B. Christensen, E. Norlinder and D. E. Ruzzante (2005). Environmental correlates of population differentiation in Atlantic herring. *Evolution* **59**: 2656-2668.

Bekkevold, D., L. A. W. Clausen, S. Mariani, C. Andre, T. B. Christensen and H. Mosegaard (2007). Divergent origins of sympatric herring population components determined using genetic mixture analysis. *Marine Ecology-Progress Series* **337**: 187-196.

Bernatchez, L. and C. Landry (2003). MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *Journal of Evolutionary Biology* **16**: 363-377.

Bierne, N. (2010). The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution* **64**: 3254-3272.

Blaxter, J. H. S. and J. R. Hunter (1982). The biology of the *Clupeoid* fishes. *Advances in Marine Biology* **20**: 3-223.

Bonin, A. (2008). Population genomics: a new generation of genome scans to bridge the gap with functional genomics. *Molecular Ecology* **17**: 3583-3584.

Bonin, A., P. Taberlet, C. Miaud and F. Pompanon (2006). Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Molecular Biology and Evolution* **23**: 773-783.

Bouck, A. and T. Vision (2007). The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology* **16**: 907-924.

Bradbury, I. R., S. Hubert, B. Higgins, T. Borza, S. Bowman, I. G. Paterson, P. V. R. Snelgrove, C. J. Morris, R. S. Gregory, D. C. Hardie, et al. (2010). Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society B-Biological Sciences* **277**: 3725-3734.

Bradshaw, W. E. and C. M. Holzapfel (2006). Climate change - Evolutionary response to rapid climate change. *Science* **312**: 1477-1478.

Carvalho, G. R. and L. Hauser (1994). Molecular-genetics and the stock concept in fisheries. *Reviews in Fish Biology and Fisheries* **4**: 326-350.

Casini, M., J. Lovgren, J. Hjelm, M. Cardinale, J. C. Molinero and G. Kornilovs (2008). Multi-level trophic cascades in a heavily exploited open marine ecosystem. *Proceedings of the Royal Society B-Biological Sciences* **275**: 1793-1801.

Clarke, B. (1975). Contribution of ecological genetics to evolutionary theory - Detecting direct effects of natural-selection on particular polymorphic loci. *Genetics* **79**: 101-113.

Colosimo, P. F., K. E. Hosemann, S. Balabhadra, G. Villarreal, M. Dickson, J. Grimwood, J. Schmutz, R. M. Myers, D. Schluter and D. M. Kingsley (2005). Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* **307**: 1928-1933.

Coop, G., D. Witonsky, A. Di Rienzo and J. K. Pritchard (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185**: 1411-1423.

Cossins, A. R. and D. L. Crawford (2005). Opinion - Fish as models for environmental genomics. *Nature Reviews Genetics* **6**: 324-333.

Dalziel, A. C., S. M. Rogers and P. M. Schulte (2009). Linking genotypes to phenotypes and fitness: how mechanistic biology can inform molecular ecology. *Molecular Ecology* **18**: 4997-5017.

Debes, P. V., F. E. Zachos and R. Hanel (2008). Mitochondrial phylogeography of the European sprat (*Sprattus sprattus* L., Clupeidae) reveals isolated climatically vulnerable populations in the Mediterranean Sea and range expansion in the northeast Atlantic. *Molecular Ecology* **17**: 3873-3888.

DeWoody, J. A. and J. C. Avise (2000). Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. *Journal of Fish Biology* **56**: 461-473.

Dulvy, N. K., S. I. Rogers, S. Jennings, V. Stelzenmuller, S. R. Dye and H. R. Skjoldal (2008). Climate change and deepening of the North Sea fish assemblage: a biotic indicator of warming seas. *Journal of Applied Ecology* **45**: 1029-1039.

EEA (2002). Biodiversity indicators for European Seas. European Environment Agency. Copenhagen

Ekblom, R. and J. Galindo (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**: 1-15.

Ellegren, H. (2008). Sequencing goes 454 and takes large-scale genomics into the wild. *Molecular Ecology* **17**: 1629-1631.

Estoup, A., P. Jarne and J. M. Cornuet (2002). Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology* **11**: 1591-1604.

Eveno, E., C. Collada, M. A. Guevara, V. Leger, A. Soto, L. Diaz, P. Leger, S. C. Gonzalez-Martinez, M. T. Cervera, C. Plomion, et al. (2008). Contrasting patterns of selection at Pinus pinaster Ait. drought stress candidate genes as revealed by genetic differentiation analyses. *Molecular Biology and Evolution* **25**: 417-437.

Ewens, W. J. (1972). Sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**: 87-&.

Excoffier, L., T. Hofer and M. Foll (2009). Detecting loci under selection in a hierarchically structured population. *Heredity* **103**: 285-298.

Excoffier, L., P. E. Smouse and J. M. Quattro (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes - Application to human mitochondrial-DNA restriction data. *Genetics* **131**: 479-491.

FAO (2010). The state of world fisheries and aquaculture. Fisheries and Aquaculture Department. Rome 2011

Faubet, P. and O. E. Gaggiotti (2008). A new Bayesian method to identify the environmental factors that influence recent migration. *Genetics* **178**: 1491-1504.

Foll, M. and O. Gaggiotti (2006). Identifying the environmental factors that determine the genetic structure of Populations. *Genetics* **174**: 875-891.

Foll, M. and O. Gaggiotti (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* **180**: 977-993.

Ford, M. J. (2002). Applications of selective neutrality tests to molecular ecology. *Molecular Ecology* **11**: 1245-1262.

Fraser, D. J., L. K. Weir, L. Bernatchez, M. M. Hansen and E. B. Taylor (2011). Extent and scale of local adaptation in salmonid fishes: review and meta-analysis. *Heredity* **106**: 404-420.

Gaggiotti, O. E., D. Bekkevold, H. B. H. Jørgensen, M. Foll, G. R. Carvalho, C. Andre and D. E. Ruzzante (2009). Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution* **63**: 2939-2951.

Galarza, J. A., J. Carreras-Carbonell, E. Macpherson, M. Pascual, S. Roques, G. F. Turner and C. Rico (2009). The influence of oceanographic fronts and early-life-history traits on

connectivity among littoral fish species. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 1473-1478.

Galtier, N., B. Nabholz, S. Glemin and G. D. D. Hurst (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Molecular Ecology* **18**: 4541-4550.

Garrigan, D. and P. W. Hedrick (2003). Perspective: Detecting adaptive molecular polymorphism: Lessons from the MHC. *Evolution* **57**: 1707-1722.

Garvin, M. R., K. Saitoh and A. J. Gharrett (2010). Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources* **10**: 915-934.

Geraldes, A., J. Pang, N. Thiessen, T. Cezard, R. Moore, Y. J. Zhao, A. Tam, S. C. Wang, M. Friedmann, I. Birol, et al. (2011). SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Molecular Ecology Resources* **11**: 81-92.

Giger, T., L. Excoffier, P. J. R. Day, A. Champigneulle, M. M. Hansen, R. Powell and C. R. Largiader (2006). Life history shapes gene expression in salmonids. *Current Biology* **16**: R281-R282.

Gomez-Uchida, D., T. W. Knight and D. E. Ruzzante (2009). Interaction of landscape and life history attributes on genetic diversity, neutral divergence and gene flow in a pristine community of salmonids. *Molecular Ecology* **18**: 4854-4869.

Grant, W. S. and F. M. Utter (1980). Biochemical genetic-variation in Walleye pollock, *Theragra-chalcogramma* - Population-structure in the Southeastern Bering Sea and the Gulf of Alaska. *Canadian Journal of Fisheries and Aquatic Sciences* **37**: 1093-1100.

Griffin, F. J., M. C. Pillai, C. A. Vines, J. Kaaria, T. Hibbard-Robbins, R. Yanagimachi and G. N. Cherr (1998). Effects of salinity on sperm motility, fertilization, and development in the Pacific herring, *Clupea pallasi*. *Biological Bulletin* **194**: 25-35.

Guinand, B., C. Lemaire and F. Bonhomme (2004). How to detect polymorphisms undergoing selection in marine fishes? A review of methods and case studies, including flatfishes. *Journal of Sea Research* **51**: 167-182.

Halverson, A. (2010). An entirely synthetic fish. New Haven & London, Yale University Press.

Hansen, M. M. and J. Hemmer-Hansen (2007). Landscape genetics goes to sea. *J Biol* **6**: 6.

Hansen, M. M., D. E. Ruzzante, E. E. Nielsen, D. Bekkevold and K. L. D. Mensberg (2002). Long-term effective population sizes, temporal stability of genetic composition and potential for local adaptation in anadromous brown trout (*Salmo trutta*) populations. *Molecular Ecology* **11**: 2523-2535.

Hanski, I. (1998). Metapopulation dynamics. *Nature* **396**: 41-49.

Hauser, L. and G. R. Carvalho (2008). Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish and Fisheries* **9**: 333-362.

Hauser, L. and J. E. Seeb (2008). Advances in molecular technology and their impact on fisheries genetics. *Fish and Fisheries* **9**: 473-486.

Hedrick, P. W. (1999). Perspective: Highly variable loci and their interpretation in evolution and conservation. *Evolution* **53**: 313-318.

Hedrick, P. W. (2005a). Genetics of populations, 3rd edn,, Jones and Bartlett.

Hedrick, P. W. (2005b). A standardized genetic differentiation measure. *Evolution* **59**: 1633-1638.

Heller, R., J. B. A. Okello and H. Siegismund (2010). Can small wildlife conservancies maintain genetically stable populations of large mammals? Evidence for increased genetic drift in geographically restricted populations of Cape buffalo in East Africa. *Molecular Ecology* **19**: 1324-1334.

Hemmer-Hansen, J., E. E. Nielsen, J. Frydenberg and V. Löeschcke (2007a). Adaptive divergence in a high gene flow environment: *Hsc70* variation in the European flounder (*Platicthys flesus L.*). *Heredity* **99**: 592-600.

Hemmer-Hansen, J., E. E. Nielsen, P. Grønkjaer and V. Löeschcke (2007b). Evolutionary mechanisms shaping the genetic population structure of marine fishes; lessons from the European flounder (*Platichthys flesus* L.). *Molecular Ecology* **16**: 3104-3118.

Hemmer-Hansen, J., E. E. G. Nielsen, D. Meldrup and C. Mittelholzer (2011). Identification of single nucleotide polymorphisms in candidate genes for growth and reproduction in a nonmodel organism; the Atlantic cod, *Gadus morhua*. *Molecular Ecology Resources* **11**: 71-80.

Hendry, A. P., T. Bohlin, B. Jonsson and O. K. Berg (2003). To sea or not to sea? Anadromy vs. non-anadromy in salmonids.Pp. 92-126 *in* A. P. Hendry and S. C. Stearns, eds. Evolution Illuminated: Salmon and Their Relatives. Oxford University Press, New York, NY.

Hendry, A. P. and S. C. Stearns, Eds. (2004). Evolution Illuminated: Salmon and Their Relatives, Oxford University Press. New York, NY.

Hermisson, J. (2009). Who believes in whole-genome scans for selection? *Heredity* **103**: 283-284.

Hewitt, G. M. (2000). The genetic legacy of the Quaternary ice ages. *Nature* **405**: 907-913.

Hohenlohe, P. A., S. J. Amish, J. M. Catchen, F. W. Allendorf and G. Luikart (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources* **11**: 117-122.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson and W. A. Cresko (2010). Population genomics of parallel adaptation in Threespine stickleback using sequenced RAD Tags. *Plos Genetics* **6**: e1000862.

Holderegger, R., U. Kamm and F. Gugerli (2006). Adaptive vs. neutral genetic diversity: implications for landscape genetics. *Landscape Ecology* **21**: 797-807.

Holderegger, R. and H. H. Wagner (2006). A brief guide to landscape genetics. *Landscape Ecology* **21**: 793-796.

Holsinger, K. E. and B. S. Weir (2009). Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics* **10**: 639-650.

Hyten, D. L., Q. J. Song, E. W. Fickus, C. V. Quigley, J. S. Lim, I. Y. Choi, E. Y. Hwang, M. Pastor-Corrales and P. B. Cregan (2010). High-throughput SNP discovery and assay development in common bean. *BMC Genomics* **11**.

ICES (2011). Report of the Working Group on the Application of Genetics in Fisheries and Mariculture (WGAGFM). ICES CM 2011/SSGHIE:13. Bangor

Iles, T. D. and M. Sinclair (1982). Atlantic herring - Stock discreteness and abundance. *Science* **215**: 627-633.

Jarne, P. and P. J. L. Lagoda (1996). Microsatellites, from molecules to populations and back. *Trends in Ecology & Evolution* **11**: 424-429.

Joost, S., M. Kalbermatten and A. Bonin (2008). Spatial analysis method(SAM): a software tool combining molecular and environmental data to identify candidate loci for selection. *Molecular Ecology Resources* **8**: 957-960.

Jørgensen, C., K. Enberg, E. S. Dunlop, R. Arlinghaus, D. S. Boukal, K. Brander, B. Ernande, A. Gardmark, F. Johnston, S. Matsumura, et al. (2007). Managing evolving fish stocks. *Science* **318**: 1247-1248.

Jørgensen, H. B. H., M. M. Hansen, D. Bekkevold, D. E. Ruzzante and V. Loeschcke (2005). Marine landscapes and population genetic structure of herring (*Clupea harengus* L.) in the Baltic Sea. *Molecular Ecology* **14**: 3219-3234.

Kawecki, T. J. and D. Ebert (2004). Conceptual issues in local adaptation. *Ecology Letters* **7**: 1225-1241.

Kelley, J. L., J. Madeoy, J. C. Calhoun, W. Swanson and J. M. Akey (2006). Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research* **16**: 980-989.

Kenchington, E. L., M. U. Patwary, E. Zouros and C. J. Bird (2006). Genetic differentiation in relation to marine landscape in a broadcast-spawning bivalve mollusc (*Placopecten magellanicus*). *Molecular Ecology* **15**: 1781-1796.

Kitano, J., S. C. Lema, J. A. Luckenbach, S. Mori, Y. Kawagishi, M. Kusakabe, P. Swanson and C. L. Peichel (2010). Adaptive divergence in the thyroid hormone signaling pathway in the stickleback radiation. *Current Biology* **20**: 2124-2130.

Klaper, R., D. Arndt, K. Setyowati, J. A. Chen and F. Goetz (2010). Functionalization impacts the effects of carbon nanotubes on the immune system of rainbow trout, *Oncorhynchus mykiss*. *Aquatic Toxicology* **100**: 211-217.

Knutsen, H., E. M. Olsen, P. E. Jorde, S. H. Espeland, C. Andre and N. C. Stenseth (2011). Are low but statistically significant levels of genetic differentiation in marine fishes 'biologically meaningful'? A case study of coastal Atlantic cod. *Molecular Ecology* **20**: 768-783.

Landry, C. R. and N. Aubin-Horth (2007). Ecological annotation of genes and genomes through ecological genomics. *Molecular Ecology* **16**: 4419-4421.

Larmuseau, M. H. D., J. A. M. Raeymaekers, K. G. Ruddick, J. K. J. Van Houdt and F. A. M. Volckaert (2009). To see in different seas: spatial variation in the rhodopsin gene of the sand goby (*Pomatoschistus minutus*). *Molecular Ecology* **18**: 4227-4239.

Larsen, P. F., E. E. Nielsen, T. D. Williams, J. Hemmer-Hansen, J. K. Chipman, M. Kruhoffer, P. Grønkjær, S. G. George, L. Dyrskjot and V. Loeschcke (2007). Adaptive differences in gene expression in European flounder (*Platichthys flesus*). *Molecular Ecology* **16**: 4674-4683.

Lepoittevin, C., J. M. Frigerio, P. Garnier-Gere, F. Salin, M. T. Cervera, B. Vornam, L. Harvengt and C. Plomion (2010). In vitro vs in silico detected SNPs for the development of a genotyping array: What can we learn from a non-model species? *PLoS ONE* **5**: e11034.

Lewontin, R. C. and J. Krakauer (1973). Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics* **74**: 175-195.

Limborg, M. T. (2007). Genetic population structure of European sprat (Sprattus sprattus L.). Master thesis. University of Copenhagen. Copenhagen

Luikart, G., P. R. England, D. Tallmon, S. Jordan and P. Taberlet (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics* **4**: 981-994.

MacCrimmon, H. R. (1971). World distribution of rainbow trout (*Salmo gairdneri*). *Journal of the Fisheries Research Board of Canada* **28**: 663-704.

Mackenzie, B. R., H. Gislason, C. Mollmann and F. W. Koster (2007). Impact of 21st century climate change on the Baltic Sea fish community and fisheries. *Global Change Biology* **13**: 1348-1367.

MacKenzie, S. A., N. Roher, S. Boltaña and F. W. Goetz (2010). Peptidoglycan, not endotoxin, is the key mediator of cytokine gene expression induced in rainbow trout macrophages by crude LPS. *Molecular Immunology* **47**: 1450-1457.

Maggs, C. A., R. Castilho, D. Foltz, C. Henzler, M. T. Jolly, J. Kelly, J. Olsen, K. E. Perez, W. Stam, R. Vainola, et al. (2008). Evaluating signatures of glacial refugia for North Atlantic benthic marine taxa. *Ecology* **89**: S108-S122.

Manel, S., M. K. Schwartz, G. Luikart and P. Taberlet (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution* **18**: 189-197.

Manni, F., E. Guerard and E. Heyer (2004). Geographic patterns of (genetic, morphologic, linguistic) variation: How barriers can be detected by using Monmonier's algorithm. *Human Biology* **76**: 173-190.

Mariani, S., W. F. Hutchinson, E. M. C. Hatfield, D. E. Ruzzante, E. J. Simmonds, T. G. Dahlgren, C. Andre, J. Brigham, E. Torstensen and G. R. Carvalho (2005). North Sea herring population structure revealed by microsatellite analysis. *Marine Ecology-Progress Series* **303**: 245-257.

Martinez, A., J. C. Garza and D. E. Pearse (2011). A microsatellite genome screen identifies chromosomal regions under differential selection in steelhead and rainbow trout. *Transactions of the American Fisheries Society* **140**: 829-842.

Martinsohn, J. T. and R. Ogden (2008). A forensic genetic approach to European fisheries enforcement. *Forensic Science International: Genetics Supplement Series* **1**: 610-611.

Martinsohn, J. T. and R. Ogden (2009). FishPopTrace—Developing SNP-based population genetic assignment methods to investigate illegal fishing. *Forensic Science International: Genetics Supplement Series* **2**: 294-296.

Maynard Smith, J. and J. Haigh (1974). The hitchhiking effect of a favorable gene. *Genetical Research* **23**: 23-35.

McCairns, R. J. S., S. Bourget and L. Bernatchez (2011). Putative causes and consequences of MHC variation within and between locally adapted stickleback demes. *Molecular Ecology* **20**: 486-502.

McCusker, M. R., E. Parkinson and E. B. Taylor (2000). Mitochondrial DNA variation in rainbow trout (*Oncorhynchus mykiss*) across its native range: testing biogeographical hypotheses and their relevance to conservation. *Molecular Ecology* **9**: 2089-2108.

Meier, K., M. M. Hansen, D. Bekkevold, O. Skaala and K. L. D. Mensberg (2011). An assessment of the spatial scale of local adaptation in brown trout (*Salmo trutta* L.): footprints of selection at microsatellite DNA loci. *Heredity* **106**: 488-499.

Meirmans, P. G. and P. W. Hedrick (2011). Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources* **11**: 5-18.

Metzker, M. L. (2010). Applications of Next-Generation Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**: 31-46.

Milano, I., M. Babbucci, F. Panitz, R. Ogden, R. O. Nielsen, M. I. Taylor, S. J. Helyar, G. R. Carvalho, M. Espiñeira, M. Atanassova, et al. (2011). Novel tools for conservation genomics: Comparing two high-throughput approaches for SNP discovery in the transcriptome of the European hake. *PLoS ONE* **6**: e28008.

Miller, D. D. and S. Mariani (2010). Smoke, mirrors, and mislabeled cod: poor transparency in the European seafood industry. *Frontiers in Ecology and the Environment* **8**: 517-521.

Miller, K. M., K. H. Kaukinen, T. D. Beacham and R. E. Withler (2001). Geographic heterogeneity in natural selection on an MHC locus in sockeye salmon. *Genetica* **111**: 237-257.

Miller, M. R., J. P. Brunelli, P. A. Wheeler, S. Liu, C. E. Rexroad, Y. Palti, C. Q. Doe and G. H. Thorgaard (2011). A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology* **doi: 10.1111/j.1365-294X.2011.05305.x**.

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko and E. A. Johnson (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* **17**: 240-248.

Moen, T., B. Hayes, F. Nilsen, M. Delghandi, K. T. Fjalestad, S. E. Fevolden, P. R. Berg and S. Lien (2008). Identification and characterisation of novel SNP markers in Atlantic cod: Evidence for directional selection. *BMC Genetics* **9**.

Morin, P. A., G. Luikart, R. K. Wayne and S. W. Grp (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution* **19**: 208-216.

Nævdal, G. (1968). Studies on hemoglobins and serum proteins in sprat from Norwegian waters. *Fiskeridirektoratets skrifter. Serie havundersøkelser* **14**: 160-182.

Naish, K. A. and J. J. Hard (2008). Bridging the gap between the genotype and the phenotype: linking genetic variation, selection and adaptation in fishes. *Fish and Fisheries* **9**: 396-422.

Namroud, M. C., J. Beaulieu, N. Juge, J. Laroche and J. Bousquet (2008). Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology* **17**: 3599-3613.

Narum, S. R., N. R. Campbell, C. C. Kozfkay and K. A. Meyer (2010). Adaptation of redband trout in desert and montane environments. *Molecular Ecology* **19**: 4622-4637.

Narum, S. R. and J. E. Hess (2011). Comparison of F_{ST} outlier tests for SNP loci under selection. *Molecular Ecology Resources* **11**: 184-194.

Narum, S. R., J. S. Zendt, C. Frederiksen, N. Campbell, A. Matala and W. R. Sharp (2011). Candidate genetic markers associated with anadromy in *Oncorhynchus mykiss* of the Klickitat River. *Transactions of the American Fisheries Society* **140**: 843-854.

Nielsen, E. E., M. M. Hansen and D. Meldrup (2006). Evidence of microsatellite hitch-hiking selection in Atlantic cod (*Gadus morhua* L.): implications for inferring population structure in nonmodel organisms. *Molecular Ecology* **15**: 3219-3229.

Nielsen, E. E., M. M. Hansen, D. E. Ruzzante, D. Meldrup and P. Grønkjær (2003). Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea revealed by individual admixture analysis. *Molecular Ecology* **12**: 1497-1508.

Nielsen, E. E., M. M. Hansen, C. Schmidt, D. Meldrup and P. Grønkjær (2001). Fisheries - Population of origin of Atlantic cod. *Nature* **413**: 272-272.

Nielsen, E. E., J. Hemmer-Hansen, P. F. Larsen and D. Bekkevold (2009a). Population genomics of marine fishes: identifying adaptive variation in space and time. *Molecular Ecology* **18**: 3128-3150.

Nielsen, E. E., J. Hemmer-Hansen, N. A. Poulsen, V. Loeschcke, T. Moen, T. Johansen, C. Mittelholzer, G. L. Taranger, R. Ogden and G. R. Carvalho (2009b). Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology* **9**: 11.

Nielsen, E. E. and E. Kenchington (2001). A new approach to prioritizing marine fish and shellfish populations for conservation. *Fish and Fisheries* **2**: 328-343.

Nielsen, E. E., P. H. Nielsen, D. Meldrup and M. M. Hansen (2004). Genetic population structure of turbot (*Scophthalmus maximus* L.) supports the presence of multiple hybrid zones for marine fishes in the transition zone between the Baltic Sea and the North Sea. *Molecular Ecology* **13**: 585-595.

Nielsen, E. E., P. J. Wright, J. Hemmer-Hansen, N. A. Poulsen, L. M. Gibb and D. Meldrup (2009c). Micro geographical population structure of cod *Gadus morhua* in the North Sea and west of Scotland: the role of sampling loci and individuals. *Marine Ecology-Progress Series* **376**: 213-225.

Nissling, A. and L. Westin (1997). Salinity requirements for successful spawning of Baltic and Belt Sea cod and the potential for cod stock interactions in the Baltic Sea. *Marine Ecology-Progress Series* **152**: 261-271.

Nosil, P., D. J. Funk and D. Ortiz-Barrientos (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology* **18**: 375-402.

Novaes, E., D. R. Drost, W. G. Farmerie, G. J. Pappas, D. Grattapaglia, R. R. Sederoff and M. Kirst (2008). High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. *BMC Genomics* **9**.

Ogden, R. (2008). Fisheries forensics: the use of DNA tools for improving compliance, traceability and enforcement in the fishing industry. *Fish and Fisheries* **9**: 462-472.

Ouborg, N. J., C. Pertoldi, V. Loeschcke, R. Bijlsma and P. W. Hedrick (2010). Conservation genetics in transition to conservation genomics. *Trends in Genetics* **26**: 177-187.

Palumbi, S. R. (1994). Genetic-divergence, reproductive isolation, and marine speciation. *Annual Review of Ecology and Systematics* **25**: 547-572.

Perry, A. L., P. J. Low, J. R. Ellis and J. D. Reynolds (2005). Climate change and distribution shifts in marine fishes. *Science* **308**: 1912-1915.

Petereit, C., H. Haslob, G. Kraus and C. Clemmesen (2008). The influence of temperature on the development of Baltic Sea sprat (*Sprattus sprattus*) eggs and yolk sac larvae. *Marine Biology* **154**: 295-306.

Pons, O. and R. J. Petit (1996). Measuring and testing genetic differentiation with ordered versus unordered alleles. *Genetics* **144**: 1237-1245.

Poulsen, N. A., J. Hemmer-Hansen, V. Loeschcke, G. R. Carvalho and E. E. Nielsen (2011). Microgeographical population structure and adaptation in Atlantic cod *Gadus morhua*: spatio-temporal insights from gene-associated DNA markers. *Marine Ecology-Progress Series* **436**: 231-243.

Quinn, T. P. (2005). The behaviour and ecology of Pacific salmon and trout. Bethesda, MD/Seattle, WA, American Fisheries Society/University of Washington Press.

Robertson, A. (1960). A theory of limits in artificial selection. *Proceedings of the Royal Society of London Series B-Biological Sciences* **153**: 235-249.

Rogers, S. M. and L. Bernatchez (2007). The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (*Coregonas sp* Salmonidae) species pairs. *Molecular Biology and Evolution* **24**: 1423-1438.

Ruzzante, D. E., S. Mariani, D. Bekkevold, C. Andre, H. Mosegaard, L. A. W. Clausen, T. G. Dahlgren, W. F. Hutchinson, E. M. C. Hatfield, E. Torstensen, et al. (2006). Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. *Proceedings of the Royal Society B-Biological Sciences* **273**: 1459-1464.

Sarropoulou, E., D. Nousdili, A. Magoulas and G. Kotoulas (2008). Linking the genomes of nonmodel teleosts through comparative genomics. *Marine Biotechnology* **10**: 227-233.

Schindler, D. E., R. Hilborn, B. Chasco, C. P. Boatright, T. P. Quinn, L. A. Rogers and M. S. Webster (2010). Population diversity and the portfolio effect in an exploited species. *Nature* **465**: 609-612.

Schlötterer, C. (2004). The evolution of molecular markers - just a matter of fashion? *Nature Reviews Genetics* **5**: 63-69.

Secombes, C. J., T. Wang and S. Bird (2011). The interleukins of fish. *Developmental and Comparative Immunology***, doi:10.1016/j.dci.2011.05.001**.

Seeb, J. E., G. Carvalho, L. Hauser, K. Naish, S. Roberts and L. W. Seeb (2011a). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources* **11**: 1-8.

Seeb, J. E., C. E. Pascal, E. D. Grau, L. W. Seeb, W. D. Templin, T. Harkins and S. B. Roberts (2011b). Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids. *Molecular Ecology Resources* **11**: 335-348.

Selkoe, K. A., C. M. Henzler and S. D. Gaines (2008). Seascape genetics and the spatial ecology of marine populations. *Fish and Fisheries* **9**: 363-377.

Shendure, J. and H. L. Ji (2008). Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135-1145.

Sick, K. (1965a). Haemoglobin polymorphism of cod in the Baltic and the Danish Belt Sea. *Hereditas* **54**: 19-48.

Sick, K. (1965b). Haemoglobin polymorphism of cod in the North Sea and the North Atlantic Ocean. *Hereditas* **54**: 49-69.

Sommer, S. (2005). The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Front Zool* **2**: 16.

Sork, V. L. and L. Waits (2010). Contributions of landscape genetics - approaches, insights, and future potential. *Molecular Ecology* **19**: 3489-3495.

Star, B., A. J. Nederbragt, S. Jentoft, U. Grimholt, M. Malmstrom, T. F. Gregers, T. B. Rounge, J. Paulsen, M. H. Solbakken, A. Sharma, et al. (2011). The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477**: 207-210.

Stinchcombe, J. R. and H. E. Hoekstra (2008). Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* **100**: 158-170.

Storz, J. F. (2005). Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* **14**: 671-688.

Teshima, K. M., G. Coop and M. Przeworski (2006). How reliable are empirical genomic scans for selective sweeps? *Genome Research* **16**: 702-712.

Thorgaard, G. H., G. S. Bailey, D. Williams, D. R. Buhler, S. L. Kaattari, S. S. Ristow, J. D. Hansen, J. R. Winton, J. L. Bartholomew, J. J. Nagler, et al. (2002). Status and opportunities for genomics research with rainbow trout. *Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology* **133**: 609-646.

Tonteri, A., A. Vasemagi, J. Lumme and C. R. Primmer (2010). Beyond MHC: signals of elevated selection pressure on Atlantic salmon (*Salmo salar*) immune-relevant loci. *Molecular Ecology* **19**: 1273-1282.

Utter, F., D. Campton, S. Grant, G. Milner, J. Seeb and L. Wishard (1980). Population structures of indigenous salmonid species of the Pacific Northwest. In: *Salmonid Ecosystems of the North Pacific* (eds McNeil WJ, Himsworth DC), pp. 285-304. *Oregon State University, Corvalis.*

Vasemägi, A. and C. R. Primmer (2005). Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology* **14**: 3623-3642.

Vitalis, R., K. Dawson and P. Boursot (2001). Interpretation of variation across marker loci as evidence of selection. *Genetics* **158**: 1811-1823.

Waples, R. S. (1998). Separating the wheat from the chaff: Patterns of genetic differentiation in high gene flow species. *Journal of Heredity* **89**: 438-450.

Waples, R. S. and O. Gaggiotti (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology* **15**: 1419-1439.

Waples, R. S., R. W. Zabel, M. D. Scheuerell and B. L. Sanderson (2008). Evolutionary responses by native species to major anthropogenic changes to their ecosystems: Pacific salmon in the Columbia River hydropower system. *Molecular Ecology* **17**: 84-96.

Ward, R. D., M. Woodwark and D. O. F. Skibinski (1994). A comparison of genetic diversity levels in marine, fresh-water, and anadromous fishes. *Journal of Fish Biology* **44**: 213-232.

Watterson, G. A. (1978). Homozygosity test of neutrality. *Genetics* **88**: 405-417.

Wenger, S. J., D. J. Isaak, C. H. Luce, H. M. Neville, K. D. Fausch, J. B. Dunham, D. C. Dauwalter, M. K. Young, M. M. Elsner, B. E. Rieman, et al. (2011). Flow regime, temperature, and biotic interactions drive differential declines of trout species under climate change. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 14175-14180.

Whitehead, P. J. P. (1985). FAO species catalogue. Clupeoid fishes of the world. An annotated and illustrated catalogue of the herrings, sardines, pilchards, sprats, anchovies and wolfherrings. Rome, FAO Fisheries Synopsis.

Chapter 2

Imprints from genetic drift and mutation imply relative divergence times across marine transition zones in a pan-European small pelagic fish (*Sprattus sprattus*)

Published in *Heredity*

Heredity (2012) 109, 96–107

www.nature.com/hdy

Heredity

ORIGINAL ARTICLE

Imprints from genetic drift and mutation imply relative divergence times across marine transition zones in a pan-European small pelagic fish (*Sprattus sprattus*)

MT Limborg[1], R Hanel[2], PV Debes[3], AK Ring[4], C André[4], CS Tsigenopoulos[5] and D Bekkevold[1]

Geographic distributions of most temperate marine fishes are affected by postglacial recolonisation events, which have left complex genetic imprints on populations of marine species. This study investigated population structure and demographic history of European sprat (*Sprattus sprattus* L.) by combining inference from both mtDNA and microsatellite genetic markers throughout the species' distribution. We compared effects from genetic drift and mutation for both genetic markers in shaping genetic differentiation across four transition zones. Microsatellite markers revealed significant isolation by distance and a complex population structure across the species' distribution (overall $\theta_{ST} = 0.038$, $P < 0.01$). Across transition zones markers indicated larger effects of genetic drift over mutations in the northern distribution of sprat contrasting a stronger relative impact of mutation in the species' southern distribution in the Mediterranean region. These results were interpreted to reflect more recent divergence times between northern populations in accordance with previous findings. This study demonstrates the usefulness of comparing inference from different markers and estimators of divergence for phylogeographic and population genetic studies in species with weak genetic structure, as is the case in many marine species.
Heredity (2012) **109,** 96–107; doi:10.1038/hdy.2012.18; published online 2 May 2012

Keywords: transition zones; genetic drift; mutation; phylogeography; marine fish; *Sprattus sprattus*

INTRODUCTION

Disentangling the evolutionary processes shaping population structure is of fundamental importance for understanding contemporary distributions of species and populations. Species distributions are in part determined by the environmental regimes within which a full life cycle can be sustained. However, environmental conditions change over time, potentially causing distributional shifts (for example, Perry *et al.*, 2005) and may lead to isolation of demes from a previously panmictic population. Despite the usual lack of physical barriers in the sea it is now generally accepted that many marine organisms show population structures deviating from a pattern of panmixia and often distance may be the only factor restricting gene flow. Indeed, many species seem to display population structures reflecting barriers to gene flow over relatively small geographic scales (for example, Ruzzante *et al.*, 1998; Bekkevold *et al.*, 2005). Such genetic discontinuities are often referred to as phylogeographic breaks (Avise, 2000) and can arise and be maintained from a multitude of processes, including climatic and glacial cycles separating previously panmictic populations or connecting populations that diverged in allopatry (Barton and Hewitt, 1985). In the northern hemisphere, for example, population structures are highly influenced by the Quaternary glaciations (Maggs *et al.*, 2008), presumably with the strongest imprint from the last glacial maximum (LGM) ~20 000 bp. Furthermore, retention of juvenile stages by local oceanographic barriers has also been suggested to halt gene flow among contiguous populations (for example, Ruzzante *et al.*, 1998). Lastly, genetic barriers may also be maintained by natural selection acting against migrants between locally adapted populations (that is, a 'tension zone' sensu Barton and Hewitt, 1985).

Throughout European waters, at least six major phylogeographic breaks or transition zones have been described for a variety of different marine taxa. These include first the area of the Aegean Archipelago and the Dardanelle Strait separating Black Sea from Mediterranean populations (Magoulas *et al.*, 1996; Nikula and Vainola, 2003). Second, gene flow barriers separating the Adriatic from other eastern Mediterranean populations have also been reported (Stefanni and Thorley, 2003; Peijnenburg *et al.*, 2006). Third, a genetic transition zone has been described in the Siculo-Tunisian strait and/or the Strait of Messina separating populations in western and eastern Mediterranean Basins (for example, Borsa *et al.*, 1997; Rolland *et al.*, 2007). Genetic transitions between the Atlantic–Mediterranean and between the Baltic–Atlantic regions are also pronounced across many taxa, with reports of clear breaks from the Strait of Gibraltar to the Almeria–Oran front (reviewed in Patarnello *et al.*, 2007) and in the Skagerrak–western Baltic (reviewed in Johannesson and Andre, 2006). Lastly, the English Channel has also

[1]Section for Population Ecology and Genetics, National Institute of Aquatic Resources, Technical University of Denmark, Silkeborg, Denmark; [2]Institute of Fisheries Ecology, Johann Heinrich von Thünen-Institut (vTI), Federal Research Institute for Rural Areas, Forestry and Fisheries, Hamburg, Germany; [3]Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada; [4]Department of Marine Ecology – Tjärnö, University of Gothenburg, Strömstad, Sweden and [5]Institute of Marine Biology and Genetics (IMBG), Hellenic Centre for Marine Research (HCMR), Heraklion, Greece
Correspondence: Dr MT Limborg, Section for Population Ecology and Genetics, National Institute of Aquatic Resources, Technical University of Denmark, Vejlsøvej 39, Silkeborg-DK-8600, Denmark.
E-mail: mol@aqua.dtu.dk
Received 10 October 2011; revised 2 March 2012; accepted 15 March 2012; published online 2 May 2012

been identified to constitute a transition zone in the polychaete *Pectinaria koreni* (Jolly *et al.*, 2005).

Few marine organisms are distributed throughout European continental waters, limiting the potential for conducting large-scale intraspecific comparisons of multiple transition zones. Phylogeographic studies have simultaneously spanned up to three of the above marine transition zones (e.g., Borsa *et al.*, 1997; Rolland *et al.*, 2007; Larmuseau *et al.*, 2009), but few have comprised all of them for the same species (Nikula and Vainola, 2003; Wilson and Veraguth, 2010).

Here, we use the European sprat (*Sprattus sprattus* L.) as a model for studying contemporary population structure and distribution in relation to known transition zones. Sprat is a locally abundant, small pelagic clupeid fish with a nearly pan-European distribution: ranging from the Black Sea, along the northern Mediterranean and Iberian coasts to the Atlantic, North Sea, Norwegian coastal waters and into the Baltic Sea. Sprat thus occupies highly heterogeneous environments. A study using a mtDNA marker suggested a complex phylogeographic history with two major clades: one representing the clade that presumably colonised northern European waters following the LGM, and a second in the eastern Mediterranean and the Black Sea with a pre- or postglacial origin (Debes *et al.*, 2008). A substructure was also evident within clades, with genetic differences within the 'western' clade between Atlantic–Baltic Sea and western Mediterranean populations, and within the 'eastern' clade between Adriatic Sea and Black Sea populations (Debes *et al.*, 2008). A recent microsatellite study further demonstrated population structure across the Baltic–Atlantic transition zone (Limborg *et al.*, 2009). In the current study, we analyse the combined data sets for mtDNA and microsatellite markers from the two above studies and also extend the previous sampling coverage. The combination of a new extensive sampling scheme with inference from both genetic markers allows us to gain insight into the underlying evolutionary mechanisms shaping population structure across multiple European transition zones. We then infer relative divergence times across transition zones, defined as 'old' (with a significant effect of mutation on genetic differentiation) vs 'recent' (with no significant effect of mutation on genetic differentiation), by contrasting effects of genetic drift and mutation for both marker types.

MTERIALS AND METHODS

Samples

Samples covered the species' distribution from its northern (Northeast Atlantic Ocean, North and Baltic Seas) to its southern (Mediterranean and Black Seas) range (Figure 1, Table 1). Sampling density, however, differed between north and south, reflecting a more continuous distribution of spawning locations in the north, compared with the south where major populations presumably are presently restricted to the Gulf of Lion, the northern Adriatic Sea and the Black Sea Basins (Debes *et al.*, 2008). Additional occurrences in estuarine areas around the Iberian Peninsula have been reported, but populations are thought to be in strong decline or even disappeared (Cabral *et al.*, 2001). Findings in the northern Aegean Sea have been reported (Deval *et al.,* 2002), but our own sampling efforts in that region have not been successful. Assumed occurrences in the Strait of Sicily have never been confirmed (O Jarboui, personal communication). Data from a total of 21 sampling stations representing 19 locations were included in the analysis (Figure 1). Of these, mtDNA variation was reported for seven locations in Debes *et al.* (2008). Microsatellite data were compiled for 17 sampling stations: 11 stations as reported by Limborg *et al.* (2009) and six additional stations extending the previous northerly dominated coverage southwards via the English Channel and the Atlantic into the Mediterranean and eastwards into the Black Sea (Figure 1). This data set effectively increased sampling coverage throughout the species distribution including the Mediterranean region. Samples were, with few exceptions, collected on spawning sites during the spawning season, which differs among populations. Temporal replicates were collected from two locations (see Table 1, Debes *et al.* (2008) and Limborg *et al.* (2009) for more details on sampling).

Populations and geographic transition zones studied

In the following, we refer to all geographic zones with observed genetic discontinuities as 'genetic transition zones', regardless of the underlying mechanisms. Of the six major cross-species transition zones defined *a priori*, the English Channel does not appear to constitute one for sprat, as samples on either side of the English Channel show spatial as well as temporal genetic homogeneity (Limborg *et al.*, 2009; Glover *et al.*, 2011). We could not address a potential transition zone between the Adriatic Sea and the eastern Mediterranean as no sprat could be obtained from the eastern Mediterranean. In the current study, we are thus obliged to use the Adriatic population for investigating genetic differentiation between the eastern Mediterranean and the Black Sea. We investigated genetic differentiation across the remaining four major transition zones separating the following regions: (i) the Baltic Sea from the Atlantic region (here, the latter includes the North Sea and English Channel, *Balt–Atl)*, (ii) the Atlantic region from the Mediterranean Sea (*Atl–WMed*), (iii) the western from the eastern Mediterranean Sea (*WMed–EMed*), and (iv) the eastern Mediterranean Sea from the Black Sea (*EMed–Black*; Figure 1).

Molecular analyses

Samples from seven locations were genotyped for both mitochondrial and nuclear DNA markers, of which both marker types were analysed for the same individuals in three of the samples (Table 1). Thus, four samples were typed only for mtDNA and 14 only for microsatellite markers.

In total, 210 individuals from seven locations (Table 1) were sequenced for a partial fragment of the 5′-end of the mitochondrial control region, as described in Debes *et al.* (2008).

A total of 1531 individuals, including 556 new to this study, were typed for nine species-specific microsatellite loci: *Spsp47D*, *Spsp77C*, *Spsp133*, *Spsp155*, *Spsp170*, *Spsp202*, *Spsp219*, *Spsp256* and *Spsp275* (Dailianis *et al.*, 2008). DNA extraction and PCR amplification were performed as described in Dailianis *et al.* (2008). PCR-amplified microsatellite fragments were analysed either on a BaseStation 51 DNA fragment analyser (MJ Research, Skovlunde, Denmark) followed by semi-automatically typing of genotypes with the software CARTOGRAPHER 1.2.6 (MJ Geneworks Inc., Skovlunde, Denmark) (samples GOT, GDA, BOR05, BOR06, ARK, BEL, KAT, SKA, GER04, GER05, ENC, CEL, BoB and ADR), or on a Beckman Coulter CEQ 8000 (Beckman-Coulter, Fullerton, CA, USA) automated sequencer (samples SKA, LIO, BLW and BLE). For the latter, allele sizes were scored with the software CEQ 8000 Genetic Analysis System (version 8.0.52; Beckman-Coulter). All individual runs included a 400-bp ladder (Applied Biosystems, Foster City, CA, USA; Beckman-Coulter). To obtain consistency in genotype scoring among runs and between platforms, we (i) analysed from two to four heterozygote control individuals spanning the anticipated allelic ranges, (ii) double-typed two samples ($n = 40$) on both platforms, and (iii) split the SKA sample into two groups of $n = 50$ and genotyped on different platforms to test for consistency in allele frequency estimates between platforms (see Supplementary File S1 for further details on validation of scoring consistency).

Genetic variation

For microsatellites, potential effects of technical or sampling artefacts were assessed by checking for effects of null alleles and departure from Hardy–Weinberg Equilibrium (HWE) and gametic phase equilibrium (LD) using MICRO-CHECKER 2.2.3 (Van Oosterhout *et al.*, 2004) and GENEPOP 4.0 (Raymond and Rousset, 1995), respectively. In all following analyses including multiple tests, results were corrected with the sequential Bonferroni method (Rice, 1989). Overall genetic variation and diversity were estimated by allelic richness (A_r) for each sample and locus using FSTAT 2.9.3 (Goudet, 1995). Weir and Cockerham′s inbreeding coefficient θ_{IS} (Weir and Cockerham, 1984) was estimated for each locus and sample using FSTAT 2.9.3. Numbers of alleles (A), expected and observed heterozygosity (H_E and H_O, respectively) were calculated for all loci and samples using Arlequin 3.5 (Excoffier and Lischer, 2010).

Figure 1 Samples analysed for microsatellites (black squares with three letter sample ID) and mtDNA (white circles with two-letter sample ID) markers. Sample ID corresponds to Table 1. Underlined samples represent locations with temporally repeated sampling. All samples analysed with the mtDNA marker (white circles) are the same as in Debes *et al.* (2008). Grey-shaded areas with labels in italics show transition zones separating the Baltic and Atlantic (including the North Sea, abbreviated *Balt–Atl*), Atlantic and Mediterranean (*Atl-WMed*), western and eastern Mediterranean (*WMed-EMed*) and the Mediterranean and Black Sea (*EMed-Black*), respectively. See text and Table 1 for more details.

Outlier analysis

Potential effects of natural or hitchhiking selection on microsatellite loci may obscure inferred patterns of neutral demographic processes (Nielsen *et al.*, 2006). We tested for any such patterns using BayeScan 1.0, following the Bayesian method described in Foll and Gaggiotti (2008). To obtain sufficient convergence of MCMC chains, we ran 10 pilot runs of 5000 iterations and an additional burn-in of 5×10^6 iterations with a thinning interval of 50 and a final sample size of 50 000. For comparison, we also used the model by Excoffier *et al.* (2009b) as implemented in Arlequin 3.5 (Excoffier and Lischer, 2010) by running 10 000 simulations.

Inference of total number of populations

To infer the number of populations in our samples we analysed the microsatellite data using the Bayesian clustering model implemented in STRUCTURE 2.3.1 (Pritchard *et al.*, 2000). This model infers population structure by clustering individual multilocus genotypes into a given number of populations (K) by minimising LD and overall departure from HWE. We used the admixture model with correlated allele frequencies among populations. We initially considered five trials for each value of K from one to ten. To ascertain adequate convergence of the MCMC model we used a burn-in of 5×10^5 iterations, followed by 2×10^6 sampled iterations. We considered the mean probability values of $\ln P(X|K)$ given by the programme, as well as the ΔK method (Evanno *et al.*, 2005) to infer the most likely number of populations. For subsequent biological interpretations of K we focused on the smallest value capturing most of the structure in the data, as suggested in the manual. Subsequently, we repeated the analysis on subsets of major clusters detected by the first run, to detect potential finer scale substructure. All analyses were performed with either no population information, or including population sample as prior information, according to Hubisz *et al.* (2009). The latter model has been shown to outperform the original model for clustering populations at weak structure (that is, F_{ST} values <0.10) and with limited numbers of microsatellite markers (Hubisz *et al.*, 2009).

Statistical analyses of overall population structure

We used Arlequin 3.5 to estimate pairwise F_{ST} from mtDNA haplotype frequencies (using conventional F-statistics based on haplotype frequencies only) between all samples, and compared these to the pairwise Φ_{ST} estimates (that is also based on genetic distances among haplotypes) reported in Debes *et al.* (2008).

Owing to a denser coverage for samples analysed with microsatellites, the description of population structure was mainly based on these markers. We thus estimated an overall and pairwise genetic differentiation using Weir and Cockerham's (1984) estimator (here, referred to as θ_{ST}) and 95% confidence intervals (CI) using the approach described in Neff and Fraser (2010). Statistical significance of pairwise θ_{ST} estimates was tested using permutation tests implemented in FSTAT 2.9.3. RSTCALC 2.2 (Goodman, 1997) was used to estimate pairwise R_{ST} between all samples and significance was tested by 1000 permutations, whereas 95% CI were obtained by bootstrapping 1000 times over loci. A principal component analysis (PCA), based on allele frequencies, for all 17 population samples was performed using PCAGEN 1.3.1 (available at: www2.unil.ch/popgen/softwares/pcagen.htm). Significance of each principal component (PC) was tested by 10 000 randomisations.

To test if the geographic pattern of genetic differentiation is caused by isolation by distance we ran Mantel tests for pairwise matrices between geographic distance and genetic distance in Arlequin 3.5 with 100 000 permutations. This was performed for both marker types and for both

Table 1 Details of sprat samples analysed including; Oceanic region, location, sample ID, marker type, sampling date, spawning condition, sample size and estimates of genetic diversity

Oceanic region	*Sample location*	*Sample ID*	*Marker*	*Regional sample groups*	*Latitude/ longitude*	*Year*	*Month*	*Mature and spawning (%)*	*No. of individuals*	*Microsatellite diversity (Ar)*	*MtDNA haplotype diversity (h ± s.d.)*
Baltic Sea	Gotland Deep	GOT	μsat	BALT	58.24° N/ 20.31° E	2006	May	100	87	15.16	—
	Gdansk Deep	GDA	μsat	BALT	54.43° N/ 18.59° E	2006	March	100	85	15.20	—
	Eastern Baltic	BA	mtDNA	BALT	55.04° N/ 18.44° E	2006	May	NA[b]	30	—	0.945 ± 0.033
	Bornholm Basin	BOR05	μsat	BALT	55.13° N/ 16.14° E	2005	April	100	80	15.25	—
		BOR06	μsat	BALT	55.34° N/ 16.25° E	2006	March	100	88	14.63	—
	Arkona Basin	ARK	μsat		55.08° N/ 13.50° E	2006	May	100	76	14.86	—
Baltic-North Sea	Belt Sea	BEL	μsat		55.42° N/ 10.25° E	2006	March	100	77	17.59	—
transition zone	Northern Kattegat	KAT	μsat		57.42° N/ 10.48° E	2006	March	100	78	18.02	—
	Eastern Skagerrak	SKA	μsat[a]		58.12° N/ 11.52° E	2008	May	100	100	19.55	—
Atlantic Ocean	North Sea	NO	mtDNA	ATLA	55.40° N/ 06.46° E	2005	July	NA[b]	30	—	0.989 ± 0.013
	German Bight	GER04	μsat	ATLA	54.15° N/ 07.12° E	2004	May	100	88	20.07	—
		GER05	μsat	ATLA	54.07° N/ 07.47° E	2005	May	100	86	19.80	—
	English Channel	ENC	μsat[a]	ATLA	51.14° N/ 01.57° E	2009	June	58	87	20.71	—
	Celtic Sea	CEL	μsat		51.59° N/ 06.46° W	2005	December	0[c]	81	18.54	—
	Bay of Biscay	BI	mtDNA	ATLA	47.18° N/ 03.16° W	2006	March	NA[b]	30	—	0.984 ± 0.016
	Bay of Biscay	BoB	μsat[a]	ATLA	47.40° N/ 02.38° W	2008	August	0[c]	91	20.33	—
Mediterranean Sea	Gulf of Lion	LIO/LI	μsat[a]/mtDNA	WMED	43.27° N/ 03.49° E	2006	July	0[c]	96/30	19.04	0.894 ± 0.045
	Adriatic Sea	ADR/AD	μsat/mtDNA	ADRI	45.36° N/ 13.34° E	2005	December	NA[b]	85/30	17.85	0.784 ± 0.078
	Bosporus	BO	mtDNA	BLAS	41.12° N/ 29.07° E	2006	June	0[c]	30	—	0.954 ± 0.027
Black Sea	Black Sea (west)	BLW/BL	μsat[a]/mtDNA	BLAS	44.37° N/ 33.50° E	2006	June	0[c]	88/30	20.12	0.920 ± 0.033
	Black Sea (east)	BLE	μsat[a]	BLAS	41.05° N/ 40.00° E	2008	December	NA[b]	94	19.36	—

Abbreviations: Ar, mean allelic richness; ATLA, Northwest Atlantic; ADRI, Adriatic Sea; BALT, Baltic Sea; BLAS, Black Sea; mtDNA, mitochondrial control region; μsat, microsatellites; WMED, western Mediterranean.
[a]Samples genotyped for the current study.
[b]Sample collected during main spawning season but maturity stage not assessed.
[c]Samples caught outside spawning season with potential inclusion of transient migrants.

measures of genetic differentiation separately (linearised equivalents of F_{ST} and Φ_{ST} for mtDNA, F_{ST} and R_{ST} for microsatellites, respectively). Geographic distance was estimated by direct shipping distance between coordinates of sampling locations calculated with the programme Netpas Distance (Netpas).

Demographic effects on population structure

Spatial population expansions are expected to result in higher population-specific F_{ST} values in marginal populations that have potentially undergone more founder events and received fewer immigrants than populations closer to an ancestral source population (Foll and Gaggiotti, 2006; Gaggiotti and Foll, 2010). To statistically test a potential effect of range expansion on population-specific differentiation, we used GESTE v2.0 (Foll and Gaggiotti, 2006) to estimate population-specific F_{ST} values following the approach by Balding and Nichols (1995). Depending on the underlying demographic history of the species, this F_{ST} estimator describes the differentiation of each population from the overall meta-population (under a migration-drift model), or from a common ancestral source population (under a fission model) (Foll and Gaggiotti, 2006).

Genetic differentiation across transition zones

Subsequent analyses focused on genetic patterns across four transition zones (Figure 1), synthesising results from the two marker types. For mtDNA data, we pooled samples fulfilling the criteria of not crossing a transition zone as well as not showing statistically significant differentiation for either the F_{ST} or Φ_{ST} pairwise estimates within the regional groups (Table 1, also see Supplementary File S2 for pairwise Φ_{ST} and F_{ST}). Applying this approach, groups of samples thus represented the following five regions: the Baltic Sea (abbreviated BALT in Table 1), the Northeast Atlantic (incl. North Sea; ATLA), the western Mediterranean (WMED), the Adriatic Sea (ADRI) and the Black Sea (incl. Strait of Bosporus; BLAS). Similarly, for microsatellite data we pooled subsets of samples showing no statistically significant pairwise θ_{ST}, to represent the same five regions (Table 1). Samples from within the Baltic–Atlantic transition zone (BEL, KAT and SKA; Figure 1) and a single sample from the Celtic Sea (CEL) that showed weak, but significant, differentiation from neighbouring samples (Supplementary File S3) were omitted from this analysis to avoid potential confounding effects from pooling non-panmictic populations.

The programme POWSIM 4.0 (Ryman and Palm, 2006) was used to evaluate statistical power of both types of markers for detecting pairwise genetic differentiation at F_{ST} levels ranging from 0.00 to 0.10. The programme simulates the divergence of two to several subpopulations from a single ancestral population through genetic drift to a given overall F_{ST} value defined by controlling effective population size (N_e) and number of generations (t). To best reflect the assumingly large N_e of sprat, we let $N_e = 10\,000$ and varied t from 0 to 2078 for simulating different levels of differentiation. After the simulation, each subpopulation was sampled at $n = 80$ and divergence from genetic homogeneity was tested with Fisher's exact test. This procedure was repeated 1000 times and the proportion of significant outcomes was used to estimate statistical power for detecting pairwise genetic differentiation. Founder events in populations of more recently colonised areas may have left a stronger imprint from genetic drift, resulting in higher levels of pairwise F_{ST} between neighbouring populations. To infer our power for detecting such events we tested four scenarios corresponding to observed genetic differentiation between populations on both sides of the four studied transition zones. Specifically, for the Baltic–Atlantic transition zone, we pooled samples from the BALT and ATLA groups (Table 1) to represent allele frequencies for the ancestral population at the onset of the simulation process. Similarly, we pooled samples for the groups flanking each of the remaining transition zones (Table 1, Figure 1). For the mtDNA analysis, we only pooled the two geographically closest samples on each side of a transition zone, as including more samples led to violation of the maximal number of alleles (or haplotypes) for a given marker (50) allowed by POWSIM, owing to a large number of private haplotypes in all samples.

Pairwise F_{ST} estimates based on mtDNA haplotype frequencies are expected to be mainly shaped through genetic drift, at least on time scales where mutations can be largely ignored. In contrast, the Φ_{ST} estimator takes the number of mutational differences among haplotypes into account and is able to reveal higher resolution on divergence time between populations having accumulated specific mutations over time (see Excoffier *et al.*, 1992 for more details). Comparisons of F_{ST} and Φ_{ST} estimates for mtDNA sequences across multiple transition zones is therefore expected to reveal relative imprints from genetic drift and mutation in explaining the level of genetic differentiation. Thus, we repeated analyses in Arlequin 3.5 using 20 000 permutations to obtain pairwise estimates of F_{ST} (using conventional F-statistics) and Φ_{ST} (using a distance matrix based on haplotype nucleotide differences corrected with the base substitution model of Tamura and Nei (1993)) between the regions represented by the five major groups described above.

The relative effects of genetic drift and mutations in explaining genetic differentiation across transition zones were also examined for microsatellite data by applying the R_{ST} permutation test in SPAGeDi 1.2 (Hardy and Vekemans, 2002). The test compares observed R_{ST} values based on allele size differences assuming a stepwise mutation model (SMM) with a corresponding frequency distribution (ρR_{ST}) obtained by randomly permuting over allelic states following an infinite allele model of mutation. A significantly larger observed R_{ST} implies a significant role of mutation for explaining population structure and suggests that divergence occurred over very long time scales (Pons and Petit, 1996; Hardy *et al.*, 2003). Significance was tested with 20 000 permutations using a one-sided test ($R_{ST} > \rho R_{ST}$) (Slatkin, 1995). Similar tests were applied for global R_{ST} estimates for each locus and all loci together. For locus *Spsp275*, a total of four individuals from the ENC and BoB samples had considerably larger alleles (50–200 bp longer) than the maximum sizes observed in all other samples. These (rare) alleles may be the results of one or more insertion events, and including them would violate the assumption of a SMM. Information for these four individuals was therefore ignored in R_{ST} permutation tests.

RESULTS

One microsatellite locus (*Spsp154*) failed to amplify consistent fragment lengths between the two genotyping platforms in the 40 calibration individuals and was discarded from further analyses. For the remaining eight loci, scoring of genotypes was consistent between the genotyping platforms (see Supplementary File S1 for more information on calibration results).

Overall genetic variation at mtDNA

For mtDNA, a total of 128 different haplotypes with 82 segregating sites were observed in the seven samples (Debes *et al.*, 2008). Haplotype diversity (h) for each sample is reported in Table 1.

Overall genetic variation at microsatellite loci

A total of 64 individuals with more than two missing genotypes were excluded, leaving 1467 individuals for which 99.97% of all loci were scored successfully (all summary statistics for each locus and sample are reported in Supplementary File S4). MICRO-CHECKER suggested the potential presence of null alleles for 35 (out of 136) sample locus pairs, and stutter-prone scoring at eight sample locus pairs. However, no general trends of a specific locus or sample were evident and subsequent analyses including or excluding information from affected loci did not change results. After correcting for multiple tests, significant deviations from HWE remained for 10 of 136 (7%) tests distributed among four loci (*Spsp275*: 3 significant tests, *Spsp219*: 1, *Spsp133*: 4 and *Spsp170*: 2) (Supplementary File S4). One out of 28 locus pairs showed significant LD (*Spsp219*, *Spsp133*). However, this was only observed in four of the 17 population samples. A similar test for LD by Limborg *et al.* (2009) on a subset of these samples did not show overall LD for any of these loci, and LD is thus not expected to incur a general bias in our analyses. Nuclear genetic diversity assessed by A_r is reported for each sample in Table 1.

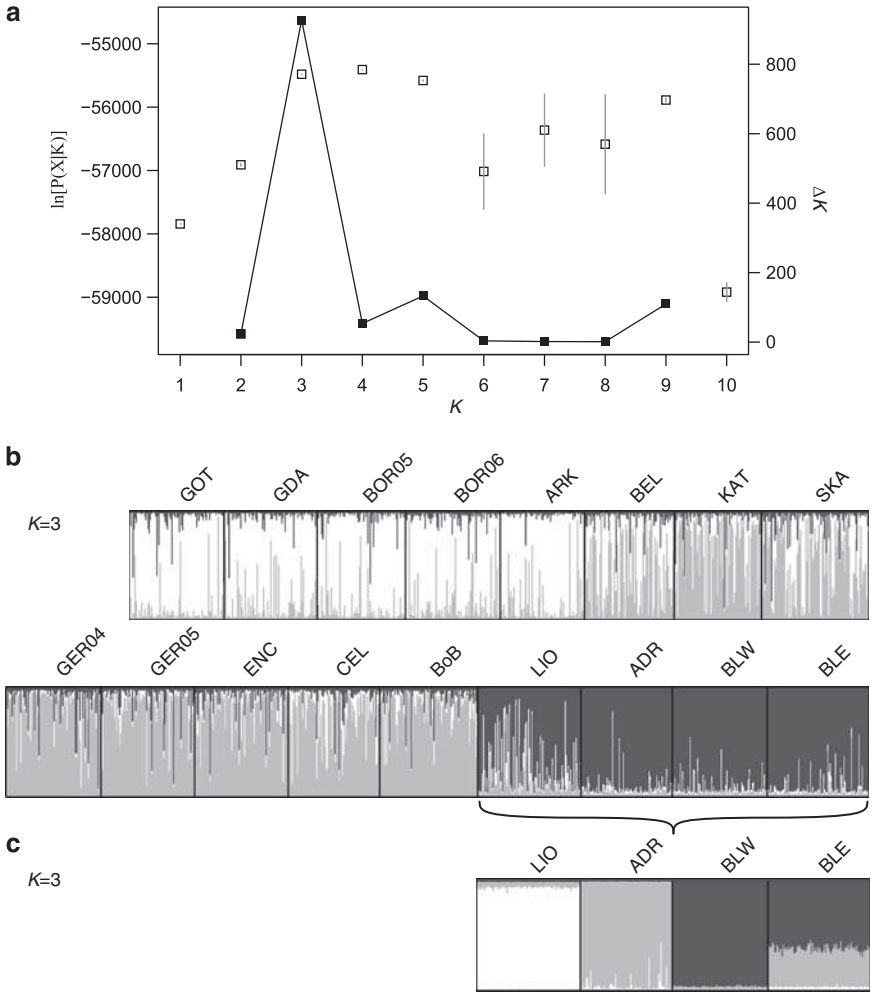


Figure 2 (**a**) Probability of each tested potential number of populations (*K*) inferred from the mean probability value ln [*P*(X|*K*)] (white squares) and the Δ*K* method (black squares) (see text for more details). (**b**) Individual population membership plotted for $K=3$. (**c**) Individual population membership when repeating the cluster analysis for the Mediterranean samples (LIO, ADR, BLW and BLE) for $K=3$ and including prior information of sample location.

Outlier analysis

The BayeScan test indicated three outlier loci (*Spsp170*, *Spsp202* & *Spsp275*) potentially subject to divergent selection, whereas the test implemented in Arlequin supported this only for the latter two loci, which also showed the highest level of support for divergent selection (Supplementary File S5). Thus, to test for potential effects of the two outlier loci found by both methods we conducted all subsequent analyses using (i) all loci, (ii) excluding each of the two outlier loci and (iii) excluding both outlier loci.

Estimation of the total number of populations

The global Bayesian clustering analysis revealed the highest likelihood for models with $K=3$ and 4, whereas the Δ*K* method suggested $K=3$ (Figure 2a) irrespective of whether prior sample information was used or not. Visual inspection revealed that setting $K>3$, did not add further meaningful inference (not shown), and we only show results for $K=3$, as this presumably captures the major biological structure across samples (that is, population clusters representing, respectively, the Baltic Sea, the Atlantic region (including the North Sea), and the Mediterranean region (including the Black Sea)) (Figure 2b). Subsequent analyses comprising either samples from within the Baltic Sea, the Atlantic region or both, with and without prior sample information did not reveal further substructuring (data not shown). When including prior sample information, an analysis comprising Mediterranean and Black Sea samples revealed substructuring ($K=3$) with the Gulf of Lion (LIO), the Adriatic (ADR) and the Black Sea (BLW and BLE) samples presumably representing genetically distinct populations (Figure 2c). The tuning parameter, *r*, for the latter model ranged from 0.06 to 0.12 among the five replicate runs. Values of *r* below 1.00 indicate that ancestry proportions differ among sampling locations and that the inclusion of prior sample information significantly increased the power for detecting weak population structure (Hubisz *et al.*, 2009). Altogether, five clusters could hence be detected using Bayesian clustering (Figure 2).

Population structure and demography

Estimators of pairwise mtDNA differentiation (F_{ST} and Φ_{ST}) revealed significant population differentiation in most comparisons (see below and Supplementary File S2). Microsatellites also revealed highly significant population structure, with an overall θ_{ST} of 0.038 (95% CI = 0.015 to 0.064, $P<0.001$) and pairwise θ_{ST} estimates ranging between 0.001 and 0.100. Genetic differentiation between temporal samples from both the Bornholm Basin (BOR) and the German Bight (GER) was low and non-significant ($\theta_{ST}<0.005$), suggesting temporal

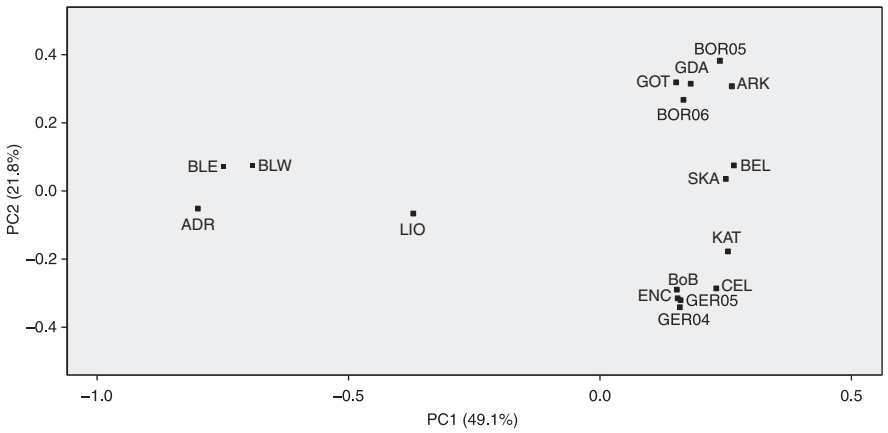


Figure 3 Genetic relationships of samples as revealed from the two first PCs from the microsatellite-based PCA. Sample IDs correspond to Figure 1.

Table 2 Results from isolation by distance tests for both marker data shown as R^2 values

Microsatellites	F_{ST}	R_{ST}
All loci	**0.82**	**0.89**
Excluding *Spsp202*	**0.72**	**0.72**
Excluding *Spsp275*	**0.66**	**0.85**
Excluding *Spsp202*&*Spsp275*	**0.30**	0.30
	F_{ST}	Φ_{ST}
mtDNA	0.01	**0.31**

Significant correlations ($\alpha = 0.05$) are shown in bold. For the microsatellite data results are also shown for tests excluding each of the two outlier loci individually or together.

stability of the observed spatial structure in these regions. The level of genetic structure varied among different geographical regions with mostly non-significant estimates of pairwise θ_{ST} within major oceanic basins in contrast to comparisons among basins (Supplementary File S3). In the PCA, the first two PCs explained a significant proportion of the total genetic variance (PC1 and PC2, $P < 0.001$; PC3 to PC10, $P = 1.000$). PC1 explained 49.1% of the total genetic variance and grouped samples corresponding to the two previously described major phylogenetic clades separated at the western and eastern Mediterranean Sea transition zone, with further separation of LIO from all other samples (Figure 3). Samples across the Baltic–Atlantic transition zone showed a clear East–West trend along PC2 (21.8%). This overall pattern remained significant although the level of differentiation was reduced when excluding the two outlier loci (not shown).

Isolation by distance was highly significant for both F_{ST} ($R^2 = 0.82$) and R_{ST} ($R^2 = 0.89$) for all microsatellite loci (Table 2). When only excluding one of the two outlier loci, results remained significant but with levels of explained variance reduced by 12–19% (Table 2). When simultaneously excluding both outlier loci the explained variation was reduced more drastically by 63–66% for both F_{ST} and R_{ST}, and only the F_{ST}-based test remained significant ($R^2 = 0.30$; Table 2). For mtDNA, a significant but weaker pattern of isolation by distance was revealed only for the Φ_{ST} values ($R^2 = 0.31$).

Population-specific F_{ST} estimates showed an increasing trend from the west (Atlantic Ocean) eastward into both the Baltic Sea in the north and into the Adriatic and Black Seas in the south (Figure 4). This overall pattern remained when excluding one or both outlier loci, although F_{ST} values reduced to 0.003–0.027 when excluding all outlier loci.

Genetic differentiation across transition zones

Overall, the power to detect genetic differentiation owing to allelic drift did not vary significantly across the four transition zones (Supplementary File S6). The mtDNA marker data lacked sufficient statistical power for detecting values of $F_{ST} < 0.02$ but could reliably detect levels of differentiation above this level ($F_{ST} = 0.02$; power = 0.872–0.952). The eight microsatellites exhibited adequate power for detecting true $F_{ST} > 0.005$ (0.998–1.00). Type-one errors ($F_{ST} = 0$) did not seem to seriously violate an assumed 5% α-level for either type of marker used (Supplementary File S6).

For the mtDNA data, pairwise F_{ST} and Φ_{ST} estimates crossing one or more of the three southern transition zones were significant for both estimators, but with considerably higher values of the Φ_{ST} estimator (Figure 5). One striking disparity, however, was observed across the northern Baltic–Atlantic transition zone where the Baltic group (BALT) showed statistically significant differentiation from the Atlantic group (ATLA) for the drift-based F_{ST} estimator but not for the Φ_{ST} estimator.

The mtDNA results were supported by the microsatellite-based analyses including all loci where the mutation-based R_{ST} estimator was not significantly higher than the purely drift-based ρR_{ST} distribution between the Baltic (BALT) and Atlantic (ATLA) groups, suggesting a negligible mutational imprint across this transition zone (Figure 6a). Also for microsatellite markers, mutations appeared to have had a relatively larger role in genetic differentiation across southern transition zones, evidenced by a significant pattern of $R_{ST} > \rho R_{ST}$ in seven comparisons (Figure 6a). A non-significant effect was observed between the Adriatic (ADRI) and Black Sea (BLAS) groups (Figure 6a). When excluding either of the two outlier loci an overall pattern of a strong mutational effect across southern transition zones remained as three and five tests remained significant when excluding *Spsp202* and *Spsp275*, respectively (Figure 6b, c). Conversely, no tests were significant when excluding both outlier loci simultaneously (Figure 6d).

DISCUSSION

By combining inference from mitochondrial and nuclear DNA markers we gained new insights into the potential effects of historical demography in explaining distribution-wide population structure of sprat covering four major transition zones. Both marker types showed clear regional patterns of population structure and especially microsatellites indicated a pattern of isolation by distance. The mtDNA marker successfully inferred old from more recent divergence times

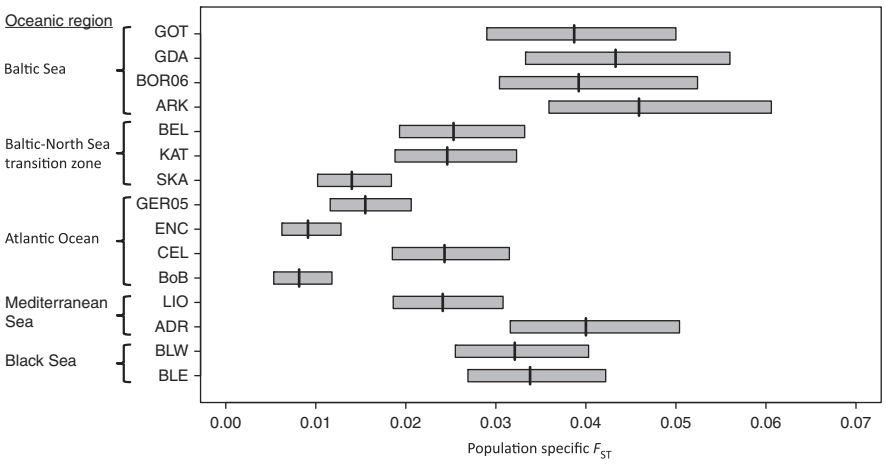


Figure 4 Population-specific F_{ST} values for microsatellite markers with black vertical bars representing mode values and grey boxes illustrating the 95% highest probability density interval (the smallest interval that contains 95% of the values). Oceanic region is given for each sample next to the vertical axis and correspond to names in Figure 1 and Table 1.

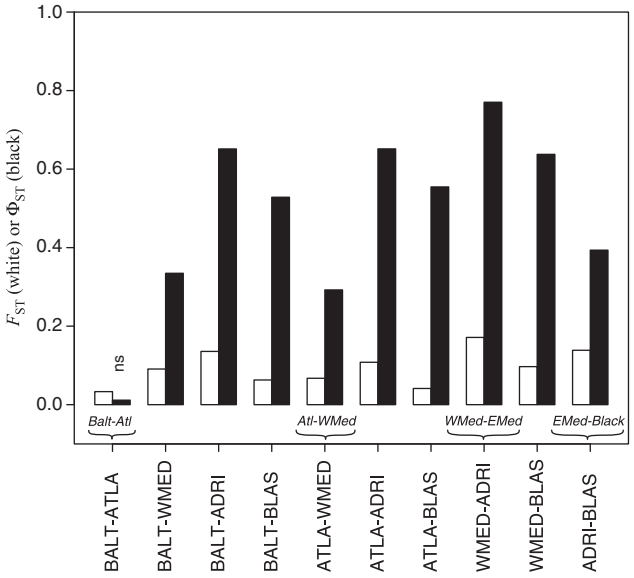


Figure 5 Pairwise F_{ST} (white bars) and Φ_{ST} (black bars) estimates for mtDNA sequences across transition zones with samples pooled into the following regions; Baltic Sea (BALT), Atlantic Ocean (ATLA), western Mediterranean (WMED), Adriatic Sea (ADRI) and the Black Sea (BLAS) (see text for more details). Pairwise comparisons between regions directly connected by each of the four transition zones are denoted with abbreviations in italic corresponding to Figure 1. All estimates are significantly >0 ($\alpha = 0.05$) unless denoted with ns.

across the different transition zones. The advantage of combining multiple marker types has previously been demonstrated in marine fishes (for example, Gonzalez and Zardoya, 2007; Wilson and Veraguth, 2010; Andre *et al.*, 2011). However, to our knowledge the present study is the first to directly compare relative imprints from genetic drift and mutation between markers and throughout the geographic distribution of a small pelagic marine fish.

Overall population structure and historical demography

The initial STRUCTURE analysis identified three population clusters and corroborated previous assertions about genetically isolated populations in the Atlantic region, the Mediterranean and Baltic Sea, respectively, using fewer samples (Limborg *et al.*, 2009). In this study, a subsequent analysis of population subsets revealed a finer structure within the Mediterranean/Black Sea region (Figure 2c). The eastern Black Sea sample (BLE) appears admixed with the Adriatic Sea population (ADR; Figure 2c), which could be explained by contemporary gene flow, shared ancestry or homoplasy. However, these explanations appear unlikely considering the intermediate location of the non-admixed sample (BLW) and this result more likely reflects analytical limitations of the method when few markers are applied (Hubisz *et al.*, 2009). Overall, the clustering result is in accordance with pairwise mtDNA and microsatellite differentiation estimates, which also revealed the highest genetic discontinuities among major oceanic basins (Supplementary Files S2 and S3) corresponding to the five clusters detected by STRUCTURE (that is, the Baltic, Atlantic, western Mediterranean, Adriatic and Black Seas).

When using STRUCTURE with the six neutrally behaving microsatellite loci only, no structure was detected (most likely $K = 1$), suggesting that the observed population structure is largely driven by the presumed outlier loci *Spsp202* and *Spsp275* (data not shown). This raises the question of whether the results indeed reflect the demographic history of the species, however, at least two facts speak in favour of this. First, the overall population structure is supported by independent analyses of both mitochondrial and nuclear DNA, and second, a PCA and estimates of pairwise θ_{ST} omitting the two outlier loci detected a similar and statistically significant (albeit weaker) pattern of population structure. Moreover, the resolving power of STRUCTURE tends to be low with few markers at low divergence (Hubisz *et al.*, 2009). Lastly, increased power is expected for detecting low genetic differentiation between predefined populations based on pairwise tests comparing allele frequencies (like θ_{ST}), compared with STRUCTURE, which does not consider such *a priori*-defined subgroups (Pritchard *et al.*, 2007).

Our results thus support a pattern with at least five more or less reproductively isolated genetic clusters in sprat throughout its distribution. Similar levels of clustering are reported for other small pelagic fishes (for example, Bekkevold *et al.*, 2005; Grant, 2005; Gonzalez and Zardoya, 2007), albeit those studies spanned narrower geographic regions. Although our focus here is the large-scale distribution, we cannot rule out the potential existence of non-sampled locally isolated populations at smaller geographic scales.

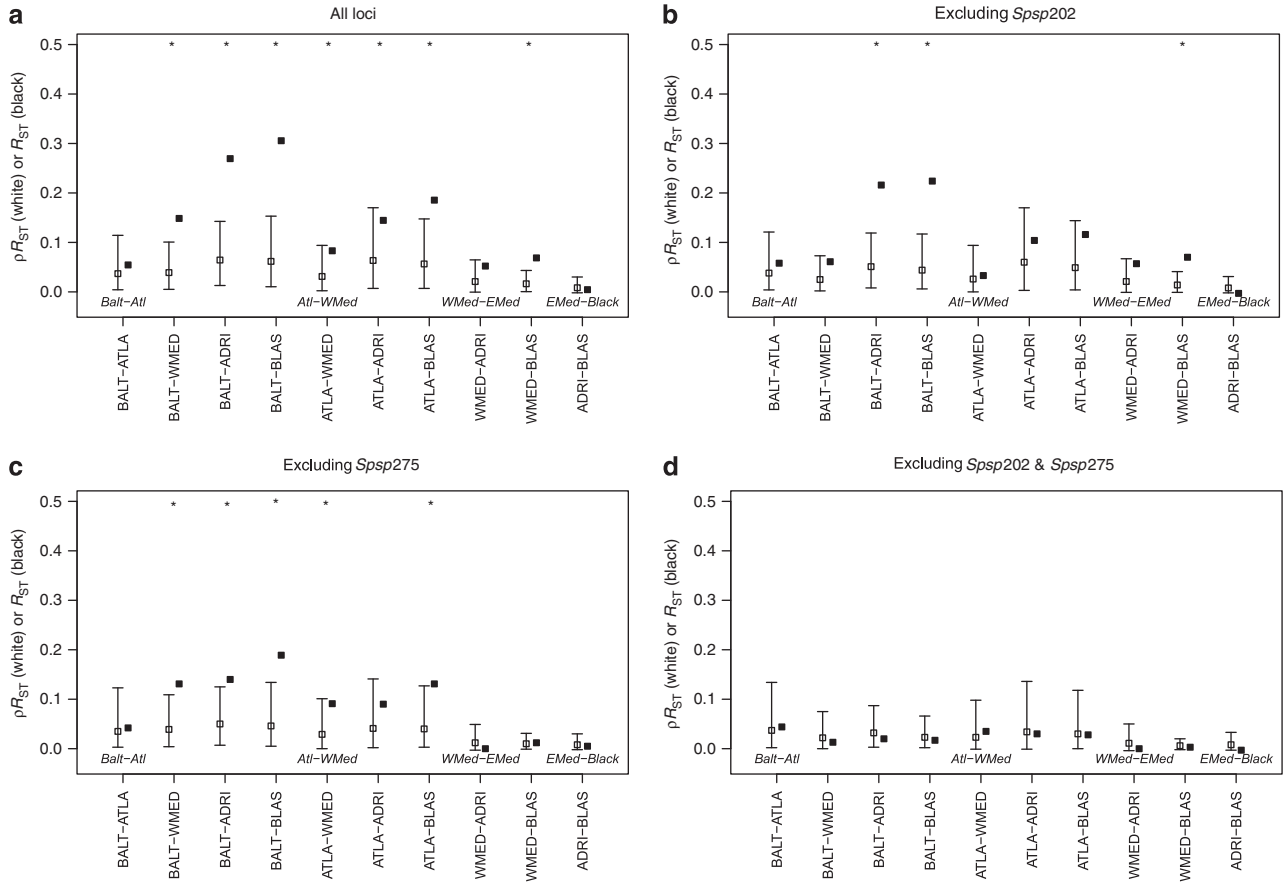


Figure 6 Observed microsatellite-based pairwise R_{ST} point estimates (black squares) between regional groups and the ρR_{ST} distribution (open squares and vertical bars represent mean and 95% CI) obtained by randomly permuting over allelic states (see text for more details). Genetic differentiations are shown across all transition zones as described for figure 5. Asterisks denote comparisons where allele size differences (that is, mutation) inferred from R_{ST} explain a significant part of the genetic differentiation. Results are shown for tests including all loci (**a**), excluding: *Spsp*202 (**b**), *Spsp*275 (**c**) and both *Spsp*202 and *Spsp*275 (**d**).

Indeed, a recent study has shown existence of population structure between Norwegian fjord populations and the North Sea sprat population (Glover *et al.*, 2011), suggesting the existence of isolated local populations.

The grouping of genetic clusters along PC1 in the PCA (Figure 3) corresponds with three distinct phylogenetic clades occurring in; (i) the Atlantic region (including the Baltic Sea), (ii) the western Mediterranean and (iii) the eastern Mediterranean (including the Black Sea). This pattern suggests that historical and phylogeographic patterns also explain a significant part of neutral genetic variation at microsatellites in combination with contemporary migration-drift processes. An effect of range expansions on genetic variation was supported by population-specific F_{ST} values which were in agreement with a 'fission model' where populations expanded from west (the Atlantic Ocean) into the Mediterranean and the Black Sea, as well as into the North and Baltic Seas. Such a demographic model was further supported by the significant patterns of isolation by distance, where especially the differentiation revealed by microsatellites was explained by geographic distance. When excluding both outlier loci, the F_{ST}-based pattern of isolation by distance was still apparent, whereas the R_{ST}-based pattern, however, became non-significant. This latter observation may indicate that for genetic differentiation at microsatellites indeed contemporary migration-drift processes may be more important than mutations, since differential mutations among populations with low or no gene flow would most likely result in larger R_{ST} estimates and significant isolation by distance. However, this result may also simply reflect technical issues if for example, neutrally behaving microsatellites are more constrained in size, which would deflate true R_{ST} values at these loci. Fragment size (alleles) distributions, however, did not suggest such a pattern in our data (not shown), thus we cannot further assess this potential explanation. Alternatively, the R_{ST} estimate may exhibit larger variance than F_{ST} (Balloux and Lugon-Moulin, 2002) explaining the observed non-significant isolation by distance pattern for R_{ST} when two outlier loci were removed.

For many marine organisms in the Northeast Atlantic, major refugia during the LGM included regions south of the Bay of Biscay with potential smaller inter-glacial refugia further north (Maggs *et al.*, 2008). For example, the thornback ray (*Raja clavata* L.) presumably persisted in at least two Atlantic refugia along the Iberian Peninsula and the Azores (Chevolot *et al.*, 2006). A similar scenario of north- and eastward expansions from one or more south westerly Atlantic refugia for sprat cannot be ruled out and would be in accordance with our results. The northwards range expansion most likely happened after the LGM in accordance with the biogeographical history of the Baltic Sea, which did not support the present-day marine fauna before ~9–7000 bp (Sohlenius *et al.*, 2001). Together with the study by Debes *et al.* (2008), our findings of large mutational differences at the

mtDNA marker suggest old population divergence across the Mediterranean transition zones potentially pre-dating the LGM. A similar scenario of pre-LGM divergence within the Mediterranean has also been suggested for another fish species (Wilson and Veraguth, 2010).

Differentiation of marginal populations

High microsatellite-based population-specific F_{ST} values and slightly reduced mtDNA haplotype diversity of Adriatic Sea and Black Sea populations compared with Atlantic samples (Figure 4, Table 1) point to a relatively old split between an eastern Mediterranean and a western Mediterranean/Atlantic clade. Debes *et al.* (2008) explained the present-day pattern at the southern edge of the distribution of sprat in the Mediterranean as a result of northwards shifting isotherms since the LGM. Populations in the northernmost Mediterranean basins occur at their physiological limit and likely represent trapped remnants of a formerly more widespread core population in the Mediterranean.

However, under this scenario, the separation in the Mediterranean of an eastern and western clade might also reflect local founder events from cryptic inter-glacial refugia pre-dating the LGM. Postglacial colonisation of the Black Sea could also, in theory, have taken place from a refugial population now only represented in this area. Moreover, the observed F_{ST} pattern (Figure 4) could also be consistent with a stepping-stone model with lower migration rates (and higher drift) for marginal populations, without inference about the directionality of founder events (Gaggiotti and Foll, 2010).

A *post hoc* permutation test in FSTAT revealed reduced allelic richness, A_r, in the Baltic group (BALT; $A_r = 15.06 \pm 0.10$ (mean ± s.e.)) compared with the Atlantic group (ATLA; $A_r = 20.23 \pm 0.14$ (mean ± s.e), one-tailed test, $P < 0.001$), consistent with observations for the mtDNA (h) (Table 1). Similar tests did not reveal significantly reduced A_r in the Adriatic Sea or Black Sea populations compared with the Atlantic group ($P > 0.11$). Assuming that the distribution of sprat populations follows a stepping-stone pattern; an alternative, but not mutually exclusive, explanation for reduced diversity and increased differentiation in the marginal Baltic Sea population can be due to reduced immigration of new alleles compared with more 'central' populations. At first sight a similar explanation appears incongruent with the relatively higher A_r in the marginal Adriatic population (Table 1). One explanation for this could be that higher microsatellite mutation rates and longer time since presumably older founder events have erased signals of reduced genetic diversity. However, strong signatures from old founder events would not be expected if contemporary immigration is the dominating factor for shaping genetic diversity in marginal populations. For example, increased environmental stress in marginal populations may reduce immigration into locally adapted populations leading to reduced diversity and greater differentiation of these populations (Excoffier *et al.*, 2009a). Lastly, congruent patterns of reduced genetic diversity in Baltic populations of other 'classical' marine fishes (reviewed in Johannesson and Andre, 2006) are suggestive of a general trend reflecting shared founder histories, reduced immigration, environmental adaptation and/or other unknown factors simultaneously reducing N_e in this marginal sea.

Disentangling effects of genetic drift and mutation across transition zones

We found indications that both genetic drift and mutation explain genetic differentiation across transition zones, but the relative effect of each varied among the different transition zones studied. This result is likely to reflect population splitting events at different time scales. More recently diverged populations will resemble each other in terms of haplotypes and alleles present, as fewer new mutations are expected to have accumulated. Pairwise differentiation between the Baltic and Atlantic groups revealed a significant F_{ST} and a lower non-significant Φ_{ST} for the mtDNA, together with a non-significant R_{ST} test for microsatellites. Genetically admixed populations within this transition zone (Figure 2) could suggest on-going gene flow eroding signals from population-specific mutations. Alternatively, recent divergence between Baltic and Atlantic populations may explain the lack of detectable differentiation in this transition zone. Although these two explanations may not be mutually exclusive, this, together with the geologic history of the Baltic region, reinforces the notion of the Baltic Sea maintaining the most recently established sprat population among those studied (see above).

Interestingly, we see a pattern of generally larger effects from mutation in most pairwise comparisons spanning one or more of the three southern transition zones. This is in accordance with the results from Debes *et al.* (2008) pointing towards relatively deep splits between samples within the Mediterranean region reflecting no or very little gene flow in combination with large divergence times. Most comparisons crossing the transition zone separating the western and eastern Mediterranean Sea, suggested by clade analysis to represent the deepest phylogeographic split (Debes *et al.*, 2008), show an accordingly larger effect from mutation and isolation. However, for microsatellites this result was mainly explained by two loci also exhibiting outlier behaviour, and thus, potentially violating the assumptions of neutrality. Furthermore, increased frequency of alleles affected by positive selection may lead to deviations from the neutral allele distribution expected under a SMM (see below). As a consequence, the mtDNA-based results may better reflect true differences between genetic drift and mutation here. The somewhat reduced mutation effect at the mtDNA between the Adriatic Sea population and the Black Sea samples (Figure 5) is in accordance with expected shorter divergence times within the two major clades and/or higher levels of gene flow (Debes *et al.*, 2008). The mtDNA-based estimators of genetic drift (F_{ST}) and mutational distance (Φ_{ST}) were both significant between the Adriatic and Black Sea groups, as opposed to across the transition zone separating the Baltic and Atlantic populations. This suggests an intermediate divergence time between the Adriatic Sea and Black Sea groups. The relatively high mutational effect between the Atlantic and western Mediterranean groups within the 'western' clade also suggest a considerably older divergence between these groups than between the Atlantic and Baltic Sea groups. However, owing to lack of Atlantic samples south of the Bay of Biscay (presumably reflecting low densities), we cannot rule out possible confounding effects from a potential undetected structure around the Iberian Peninsula. In such a case, a sample from the more southern population would be more appropriate when testing differentiation across the Atlantic–western Mediterranean transition zone owing to a potentially more recent shared ancestry with the western Mediterranean population.

A large effect of mutation relative to drift, in combination with restricted gene flow, should lead to a genome-wide pattern of $R_{ST} > \rho R_{ST}$. The significant contribution of mutations in explaining differentiations among the major clades with microsatellites was mainly driven by the two loci *Spsp202* and *Spsp275* (Supplementary File S7), which were also suggested to be affected by directional selection (Supplementary File S5). Thus, great caution should be taken when quantitatively inferring effects from mutation and neutral genetic drift when using these loci. Nevertheless, the congruent results observed for two independent loci may suggest a biological

meaningful pattern of increased mutational effects on microsatellite variation across southern transition zones. We, however, cannot rule out that these outlier loci behave in a non-neutral fashion and thus violate the SMM model leaving the microsatellite-based results inconclusive. Alternatively, our results may have indicated a general trend of microsatellite loci mainly reflecting more recent migration-drift processes as suggested from significant patterns of isolation by distance, whereas genetic variation at mtDNA markers appeared better suited for inferring older demographic histories. Indeed, for the mtDNA results we did find varying relative effects of genetic drift and mutation across different transition zones indicative of varying divergence times between different sprat populations.

Other statistical methods offer more direct estimates of bottlenecks, time since divergence and gene flow. However, the signal in our data is likely too weak to obtain reliable estimates from these analyses, as testified by the fact that analyses using the approaches of Piry *et al.* (1999) and Garza & Williamson (2001) were inconclusive of past bottlenecks in any of our populations (not shown). Also, attempts to apply IMa (Hey and Nielsen, 2004) on mtDNA and microsatellite data to infer divergence times and gene flow resulted in non-converging MCMC chains, reflecting a lack of information in the data. Instead, by taking advantage of a more indirect approach of comparing relative imprints from mutation and genetic drift, we were able to distinguish putatively 'old' from more 'recent' population divergence across transition zones putatively characterised by varying levels of gene flow. We thus expect that this approach may be useful in other applications for organisms characterised by weak structure due to recent divergence, large N_e and/or high gene flow.

DATA ARCHIVING

Data have been deposited at Dryad: doi:10.5061/dryad.m247bg66.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This study is a contribution of the EU Network of Excellence MARBEF, Marine Biodiversity and Ecosystem Functioning. The study was carried out with financial support from the European Commission through the FP6 projects UNCOVER (contract no. 022717) and RECLAIM (contract no. 044133). It does not necessarily reflect the views of the European Commission and in no way anticipates the Commission′s future policy in this area. Financial support was also partly provided by the BaltGene project funded by BONUS Baltic Organisations' Network for Funding Science EEIG, and the Linnaeus Centre for Marine Evolutionary Biology at University of Gothenburg (www.cemeb.science.gu.se). Thomas Damm Als is thanked for help with illustrations. Two anonymous referees provided constructive comments and suggestions on an earlier draft. Lastly, we are deeply grateful to Michele Casini, Georgiy Shulman, Atilla Özdemir, Melek Işinibilir, Pascal Lorance, Arnaud Souplet, Andrea Ramšak, Rudi Voss and Yves Verin for invaluable help with providing samples.

Andre C, Larsson LC, Laikre L, Bekkevold D, Brigham J, Carvalho GR *et al.* (2011). Detecting population structure in a high gene-flow species, Atlantic herring (*Clupea harengus*): direct, simultaneous evaluation of neutral vs putatively selected loci. *Heredity* **106**: 270–280.

Avise JC (2000). Phylogeography: The History and Formation of SpeciesHarvard University Press: Cambridge, MA.

Balding DJ, Nichols RA (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**: 3–12.

Balloux F, Lugon-Moulin N (2002). The estimation of population differentiation with microsatellite markers. *Mol Ecol* **11**: 155–165.

Barton NH, Hewitt GM (1985). Analysis of Hybrid Zones. *Annu Rev Ecol Syst* **16**: 113–148.

Bekkevold D, Andre C, Dahlgren TG, Clausen LAW, Torstensen E, Mosegaard H *et al.* (2005). Environmental correlates of population differentiation in Atlantic herring. *Evolution* **59**: 2656–2668.

Borsa P, Blanquer A, Berrebi P (1997). Genetic structure of the flounders *Platichthys flesus* and *P. stellatus* at different geographic scales. *Mar Biol* **129**: 233–246.

Cabral HN, Costa MJ, Salgado JP (2001). Does the Tagus estuary fish community reflect environmental changes? *Clim Res* **18**: 119–126.

Chevolot M, Hoarau G, Rijnsdorp AD, Stam WT, Olsen JL (2006). Phylogeography and population structure of thornback rays (*Raja clavata* L., Rajidae). *Mol Ecol* **15**: 3693–3705.

Dailianis T, Limborg M, Hanel R, Bekkevold D, Lagnel J, Magoulas A *et al.* (2008). Characterization of nine polymorphic microsatellite markers in sprat (*Sprattus sprattus* L.). *Mol Ecol Resour* **8**: 861–863.

Debes PV, Zachos FE, Hanel R (2008). Mitochondrial phylogeography of the European sprat (*Sprattus sprattus* L., Clupeidae) reveals isolated climatically vulnerable populations in the Mediterranean Sea and range expansion in the northeast Atlantic. *Mol Ecol* **17**: 3873–3888.

Deval MC, Ates C, Bök T, Oray IK (2002). Investigations of the distribution of the eggs and larvae of the sprata *Sprattus sprattus* L. 1758, in the Sea of Marmara. *ICES CM 2002/L:27* (www.Ices.dk).

Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* **14**: 2611–2620.

Excoffier L, Foll M, Petit RJ (2009a). Genetic consequences of range expansions. *Annu Rev Ecol Evol S* **40**: 481–501.

Excoffier L, Hofer T, Foll M (2009b). Detecting loci under selection in a hierarchically structured population. *Heredity* **103**: 285–298.

Excoffier L, Lischer HEL (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **10**: 564–567.

Excoffier L, Smouse PE, Quattro JM (1992). Analysis of molecular variance inferred from metric distances among dna haplotypes - application to human mitochondrial-DNA restriction data. *Genetics* **131**: 479–491.

Foll M, Gaggiotti O (2006). Identifying the environmental factors that determine the genetic structure of Populations. *Genetics* **174**: 875–891.

Foll M, Gaggiotti O (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genetics* **180**: 977–993.

Gaggiotti OE, Foll M (2010). Quantifying population structure using the F-model. *Mol Ecol Resour* **10**: 821–830.

Garza JC, Williamson EG (2001). Detection of reduction in population size using data from microsatellite loci. *Mol Ecol* **10**: 305–318.

Glover K, Skaala Ø, Limborg M, Kvamme C, Torstensen E (2011). Microsatellite DNA reveals population genetic differentiation among sprat (*Sprattus sprattus*) sampled throughout the Northeast Atlantic, including Norwegian fjords. *ICES J Mar Sci* **68**: 2145–2151.

Gonzalez EG, Zardoya R (2007). Relative role of life-history traits and historical factors in shaping genetic population structure of sardines (*Sardina pilchardus*). *BMC Evol Biol* **7**: 197.

Goodman SJ (1997). Rst Calc: a collection of computer programs for calculating estimates of genetic differentition from microsatellite data and determining their significance. *Mol Ecol* **6**: 881–885.

Goudet J (1995). FSTAT (version 1.2): a computer program to calculate F-statistics. *J Hered* **86**: 485–486.

Grant WS (2005). A second look at mitochondrial DNA variability in European anchovy (*Engraulis encrasicolus*): assessing models of population structure and the Black Sea isolation hypothesis. *Genetica* **125**: 293–309.

Hardy OJ, Charbonnel N, Freville H, Heuertz M (2003). Microsatellite allele sizes: A simple test to assess their significance on genetic differentiation. *Genetics* **163**: 1467–1482.

Hardy OJ, Vekemans X (2002). SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* **2**: 618–620.

Hey J, Nielsen R (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**: 747–760.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009). Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour* **9**: 1322–1332.

Johannesson K, Andre C (2006). Life on the margin: genetic isolation and diversity loss in a peripheral marine ecosystem, the Baltic Sea. *Mol Ecol* **15**: 2013–2029.

Jolly MT, Jollivet D, Gentil F, Thiebaut E, Viard F (2005). Sharp genetic break between Atlantic and English Channel populations of the polychaete *Pectinaria koreni*, along the North coast of France. *Heredity* **94**: 23–32.

Larmuseau MHD, Van Houdt JKJ, Guelinckx J, Hellemans B, Volckaert FAM (2009). Distributional and demographic consequences of Pleistocene climate fluctuations for a marine demersal fish in the north-eastern Atlantic. *J Biogeogr* **36**: 1138–1151.

Limborg MT, Pedersen JS, Hemmer-Hansen J, Tomkiewicz J, Bekkevold D (2009). Genetic population structure of European sprat *Sprattus sprattus*: differentiation across a steep environmental gradient in a small pelagic fish. *Mar Ecol Prog Ser* **379**: 213–224.

Maggs CA, Castilho R, Foltz D, Henzler C, Jolly MT, Kelly J *et al.* (2008). Evaluating signatures of Glacial Refugia for North Atlantic Benthic Marine Taxa. *Ecology* **89**: S108–S122.

Magoulas A, Tsimenides N, Zouros E (1996). Mitochondrial DNA phylogeny and the reconstruction of the population history of a species: the case of the European anchovy (*Engraulis encrasicolus*). *Mol Biol Evol* **13**: 178–190.

Neff BD, Fraser BA (2010). A program to compare genetic differentiation statistics across loci using resampling of individuals and loci. *Mol Ecol Resour* **10**: 546–550.

Nielsen EE, Hansen MM, Meldrup D (2006). Evidence of microsatellite hitch-hiking selection in Atlantic cod (*Gadus morhua* L.): implications for inferring population structure in nonmodel organisms. *Mol Ecol* **15**: 3219–3229.

Nikula R, Vainola R (2003). Phylogeography of *Cerastoderma glaucum* (Bivalvia: Cardiidae) across Europe: a major break in the Eastern Mediterranean. *Mar Biol* **143**: 339–350.

Patarnello T, FAMJ Volckaert, Castilho R (2007). Pillars of Hercules: is the Atlantic-Mediterranean transition a phylogeographical break? *Mol Ecol* **16**: 4426–4444.

Peijnenburg KTCA Fauvelot C, Breeuwer AJ, Menken SBJ (2006). Spatial and temporal genetic structure of the planktonic Sagitta setosa (Chaetognatha) in European seas as revealed by mitochondrial and nuclear DNA markers. *Mol Ecol* **15**: 3319–3338.

Perry AL, Low PJ, Ellis JR, Reynolds JD (2005). Climate change and distribution shifts in marine fishes. *Science* **308**: 1912–1915.

Piry S, Luikart G, Cornuet JM (1992). BOTTLENECK: A computer program for detecting recent reductions in the effective population size using allele frequency data. *J Hered* **90**: 502–503.

Pons O, Petit RJ (1996). Measuring and testing genetic differentiation with ordered versus unordered alleles. *Genetics* **144**: 1237–1245.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Pritchard J K, Wen X, Falush D (2007). *Documentation for structure software: Version 2.2*. Available from http://pritch.bsd.uchicago.edu.

Raymond M, Rousset F (1995). Genepop (Version-1.2) - population-genetics software for exact tests and ecumenicism. *J Hered* **86**: 248–249.

Rice WR (1989). Analyzing tables of statistical tests. *Evolution* **43**: 223–225.

Rolland JL, Bonhomme F, Lagardere F, Hassan M, Guinand B (2007). Population structure of the common sole (*Solea solea*) in the Northeastern Atlantic and the Mediterranean Sea: revisiting the divide with EPIC markers. *Mar Biol* **151**: 327–341.

Ruzzante DE, Taggart CT, Cook D (1998). A nuclear DNA basis for shelf- and bank-scale population structure in northwest Atlantic cod (*Gadus morhua*): Labrador to Georges Bank. *Mol Ecol* **7**: 1663–1680.

Ryman N, Palm S (2006). POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Mol Ecol Notes* **6**: 600–602.

Slatkin M (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**: 457–462.

Sohlenius G, Emeis KC, Andren E, Adren T, Kohly A (2001). Development of anoxia during the Holocene fresh-brackish water transition in the Baltic Sea. *Mar Geol* **177**: 221–242.

Stefanni S, Thorley JL (2003). Mitochondrial DNA phylogeography reveals the existence of an Evolutionarily Significant Unit of the sand goby *Pomatoschistus minutus* in the Adriatic (Eastern Mediterranean). *Mol Phylogenet Evol* **28**: 601–609.

Tamura K, Nei M (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. *Mol Biol Evol* **10**: 512–526.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004). MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Mol Ecol Notes* **4**: 535–538.

Weir BS, Cockerham CC (1984). Estimating F-Statistics for the analysis of population-structure. *Evolution* **38**: 1358–1370.

Wilson AB, Veraguth IE (2010). The impact of Pleistocene glaciation across the range of a widespread European coastal species. *Mol Ecol* **19**: 4535–4553.

Supplementary material accompanies the paper on Heredity website (http://www.nature.com/hdy)

Chapter 3

Genetic population structure of European sprat (*Sprattus sprattus* L.): differentiation across a steep environmental gradient in a small pelagic fish

Published in *Marine Ecology-Progress series*

Vol. 379: 213–224, 2009
doi: 10.3354/meps07889

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published March 30

Genetic population structure of European sprat *Sprattus sprattus*: differentiation across a steep environmental gradient in a small pelagic fish

Morten T. Limborg[1,2,*], Jes S. Pedersen[1], Jakob Hemmer-Hansen[2], Jonna Tomkiewicz[3], Dorte Bekkevold[2]

[1]University of Copenhagen, Department of Biology, Centre of Social Evolution, Universitetsparken 15, 2100 Copenhagen Ø, Denmark

[2]Technical University of Denmark, National Institute of Aquatic Resources, Section for Population Genetics, Vejlsøvej 39, 8600 Silkeborg, Denmark

[3]Technical University of Denmark, National Institute of Aquatic Resources, Section for Population and Ecosystem Dynamics, Kavalergården 6, 2920 Charlottenlund, Denmark

ABSTRACT: Factors such as oceanographic retention, isolation by distance and secondary contact zones have, among others, been suggested to explain the low, but statistically significant, neutral population structure observed in many marine fishes. European sprat *Sprattus sprattus* L. is not known to display philopatric spawning behaviour or to exhibit local retention of eggs and larvae. It thus constitutes a good model for studying population structure in a characteristic small pelagic fish with high dispersal potential and an opportunistic life history. We analysed 931 specimens of sprat from 9 spawning locations in and around the North Sea and Baltic Sea area and from a geographically distant population from the Adriatic Sea. Analyses of 9 microsatellite loci revealed a sharp genetic division separating samples from the northeastern Atlantic Ocean and the Baltic Sea (pairwise $\theta = 0.019$ to 0.035), concurring with a steep salinity gradient. We found, at most, weak structure among samples within the northeastern Atlantic region and within the Baltic Sea (pairwise $\theta = 0.001$ to 0.009). The Adriatic Sea population was highly differentiated from all northern samples (pairwise $\theta = 0.071$ to 0.092). Overall, the observed population structure resembles that of most other marine fishes studied in the North and Baltic Sea areas. Nevertheless, spatially explicit differences are observed among species, probably reflecting specific life histories. Such fine-scale population structures should be taken into account when considering complex ecosystem functions, e.g. in multispecies stock management.

KEY WORDS: European sprat · Population structure · Environmental gradients · Interspecific comparison · Salinity · Marine fishes · Microsatellite DNA

INTRODUCTION

Over the last decades ample evidence of significant, albeit commonly low, levels of genetic population differentiation has been accumulated for marine fishes (e.g. Ruzzante et al. 1998, Pampoulie et al. 2004, Jørgensen et al. 2005, Hemmer-Hansen et al. 2007b). These studies have challenged the long-held view of predominantly limited population structure in marine fishes inhabiting large coherent environments with few physical barriers. Different explanations have been proposed to account for observed population structure in marine fishes. For example, physical forcing by current systems and local gyres may retain eggs and larvae in local nursery areas (Ruzzante et al. 1998), and historical events (e.g. geological processes) can lead to genetic divergence of populations by isolating contingents of populations in temporary refugia

*Email: mol@aqua.dtu.dk

(Hewitt 2004, Knowles & Richards 2005). Furthermore, adaptation to local environments can lead to establishment of gene-flow barriers across environmental transition zones through hybrid inferiority (Barton & Hewitt 1985).

The North Sea–Baltic Sea transition zone represents a major environmental gradient, characterised by a dramatic change in salinity over a few hundred kilometres from oceanic conditions (30 to 35‰) in the Skagerrak to an average salinity of 8 to 10‰ in the Western Baltic Sea. The colonisation by marine species in the Baltic Sea is believed to have been achieved as a result of specific adaptations to life in a marginal environment (e.g. Ojaveer & Kalejs 2005). In the North Sea–Baltic Sea transition zone salinity levels are expected to exert a significant selective pressure on local populations although other environmental factors, such as temperature dynamics, are also expected to play a role. Indeed, it has been shown that Baltic (Atlantic) cod *Gadus morhua* L. tolerate lower salinities during egg fertilisation and the egg phase compared with populations from the Skagerrak (Nissling & Westin 1997). Timing of spawning also seems to conform to spatial and temporal production peaks (Tomkiewicz et al. 1998, Ojaveer & Kalejs 2005).

Molecular studies have identified genetically distinct North Sea and Baltic Sea fish populations in, for example, Atlantic herring *Clupea harengus* L. (Bekkevold et al. 2005), turbot *Psetta maxima* L. (Nielsen et al. 2004), Atlantic cod *Gadus morhua* L. (Nielsen et al. 2003) and European flounder *Platichthys flesus* L. (Hemmer-Hansen et al. 2007b), as well as in many other organisms such as algae and invertebrates (see Johannesson & Andre 2006 for a review). Overall, studies suggest restricted gene flow across the North Sea–Baltic Sea transition zone, but spatial patterns vary among species. Thus, interspecific comparisons may reveal the relative importance of specific environmental factors and/or biological traits for shaping patterns of population structure (Patarnello et al. 2007).

European sprat *Sprattus sprattus* L. is a pelagic schooling clupeid fish. Tolerating temperatures down to ~5°C (Nissling 2004) and salinities down to ~4‰ (Whitehead 1985), this species has successfully colonised a wide range of environments. Sprat is distributed in the Atlantic Ocean from the Norwegian west coast in the north to Morocco in south, including the Baltic Sea, and in the northern Mediterranean Sea and the Black Sea (Whitehead 1985). In the northeast Atlantic, spawning sprat concentrate in the deep basins of the Baltic Sea, in the Skagerrak/Kattegat area, the southeastern North Sea (German Bight) and from the English Channel to the north along the British west coast (Parmanne et al. 1994 and references therein, ICES 2007). Spawning, however, occurs throughout the species' distribution, and philopatric spawning migrations have not been described (Köster et al. 2003b). Further, local abundance and interannual movement among feeding areas can show substantial variation (Stepputtis 2006). In comparison, other species, such as cod and herring inhabiting the same areas exhibit spawning characteristics including homing and local retention of eggs that may induce stronger genetic isolation among components (Voipio 1981, Iles & Sinclair 1982, Aro 1989). In sprat, the continuous distribution of spawning habitat coupled with opportunistic vagrant behaviour (De Silva 1973, Alheit 1988) suggest limited barriers to gene flow among areas and lead to expectations of weak population differentiation in comparison with, for example, herring and cod. Based on mtDNA data, sprat has been divided into 2 major phylogenetic clades geographically separated by the Strait of Sicily, and 1 clade showing signs of a more recent (since 13 000 to 7600 yr BP) northwards expansion into the North and Baltic seas from an Atlantic refugia (Debes et al. 2008). Within the Baltic Sea, differences in meristic and morphometric characters, otolith structure and area specific stock dynamics have led several authors to suggest the occurrence of reproductively isolated populations (e.g. Aro 1989, Ojaveer 1989). However, these hypotheses have not been evaluated using genetic markers with sufficient resolution for identifying small-scale population structure (but see Kozlovski 1988).

In the present study, highly variable microsatellite markers were used to analyse European sprat samples from major spawning areas ranging from the central Baltic Sea to the Celtic Sea in the northeastern Atlantic Ocean. We ask the following questions: (1) Does sprat exhibit population structure at large (among seas) as well as regional (within sea) scales? (2) Are potential barriers to gene flow concurrent with salinity gradients in the area? (3) How does sprat population structure compare with that of other fishes in the North Sea–Baltic Sea transition zone? (4) Which biological and physical factors are likely to explain differences and similarities among species?

MATERIALS AND METHODS

Sample collection. A total of 969 sprat were collected during peak spawning time (March to May) in major spawning areas in and around the North Sea and the Baltic Sea (Table 1, Fig. 1). In total, 9 locations were sampled, of which 2 (German Bight and Bornholm Basin) included temporal replicates to test for temporal stability of genetic composition within locations. The stage of maturity was determined for all specimens, except for the Adriatic Sea sample. Prefer-

Fig. 1. *Sprattus sprattus*. Study area with sample locations. Circled numbers refer to the respective sample locations listed in Table 1

ably, only specimens in spawning condition were included in the genetic analyses to ensure proper representation of the local spawning populations. The Adriatic Sea specimens were caught in a local spawning area during spawning season. The Celtic Sea sample was collected outside the spawning season and could potentially include migrants from other populations.

Molecular analyses. DNA was extracted from fin or muscle tissue and stored in 96% ethanol using the DNeasy kit 250 (QIAGEN). Genetic variation was analysed at 9 fluorescently labelled dinucleotide microsatellite loci developed for sprat: *Spsp47D* (TET), *Spsp77C* (HEX), *Spsp133* (FAM), *Spsp154* (TET), *Spsp170* (FAM), *Spsp202* (HEX), *Spsp219* (HEX), *Spsp256* (TET) and *Spsp275* (FAM) (Dailianis et al. 2008). The loci were amplified separately by PCR using standard reagents. Annealing temperatures ranged from 56 to 62°C among loci (details in Dailianis et al. 2008). PCR amplified microsatellite fragments were analysed on a BaseStation 51™ DNA fragment analyser (MJ Research) and gels were semi-automatically typed using the software CARTOGRAPHER 1.2.6 (MJ Geneworks). Depending on marker, between 10 and 50% of the individuals from each sample were reanalysed to ensure consistency of results. Further actions were taken to minimise genotyping errors, as suggested by Bonin et al. (2004). Thus, quality of PCR products was tested on a 6% agarose gel with a negative control to rule out contamination. Further, 4 controls of known genotypes were re-run on every gel to ensure consistent scoring of genotypes. Finally, all individuals in 3 samples were cross-typed by 2 persons independently, and a third sample was typed twice by the same person (typings separated by months). Fish with 4 or more missing single-locus genotypes were omitted from the dataset.

Table 1. *Sprattus sprattus*. Location and details of sprat samples collected. Also given are percentages of spawning fish per sample and mean allelic richness (A_r) corrected for the minimum sample size (n = 56) of all loci per sample

Geographic location	Sample ID	Latitude, longitude	Year	Month	Proportion mature and spawning (%)[a]	No. of ind.	A_r
(1) Gotland Deep	GOT	58.24° N, 20.31° E	2006	May	100	88	14.3
(2) Gdansk Deep	GDA	54.43° N, 18.60° E	2006	Mar	100	86	14.1
(3) Bornholm Basin	BOR05	55.13° N, 16.14° E	2005	Apr	100	82	14.3
	BOR06	55.34° N, 16.25° E	2006	Mar	100	88	13.8
(4) Arkona Basin	ARK	55.08° N, 13.50° E	2006	May	100	78	14.1
(5) Belt Sea	BEL	55.42° N, 10.25° E	2006	Mar	100	83	16.2
(6) Northern Kattegat	KAT	57.42° N, 10.48° E	2006	Mar	100	81	16.7
(7) German Bight	GER04	54.15° N, 07.12° E	2004	May	100	88	18.5
	GER05	54.07° N, 07.47° E	2005	May	100	87	18.1
(8) Celtic Sea	CEL	51.59° N, 06.46° W	2005	Dec	0[b]	85	16.6
(9) Adriatic Sea	ADR	45.36° N, 13.34° E	2005	Dec	na[c]	85	16.2

[a]Fish in spawning phase alternating between actively spawning and final maturation of batches
[b]Caught outside spawning season
[c]Sample collected during main spawning season but maturity stage not assessed

Statistical analyses. The program MICRO-CHECKER 2.2.3 (Van Oosterhout et al. 2004) was used to test for technical artefacts, such as null alleles. Departure from Hardy-Weinberg equilibrium (HWE) was tested for each locus and sample using the method by Guo & Thompson (1992) implemented in GENEPOP 3.4 (Raymond & Rousset 1995). Analyses for departure from gametic phase equilibrium (linkage disequilibrium) between pairs of loci by means of exact tests were also performed using GENEPOP 3.4.

Observed and expected heterozygosities (H_O and H_E), Weir & Cockerham's (1984) inbreeding coefficient (F_{IS}), numbers of alleles (A) and allelic richness corrected for sample size (A_r) were calculated for each locus and sample using FSTAT 2.9.3 (Goudet 1995). Differences in allelic richness between the Baltic Sea samples: Gotland, Gdansk, Bornholm and Arkona (GOT, GDA, BOR and ARK, respectively), and samples from the northern Kattegat, German Bight and Celtic Sea (KAT, GER and CEL, respectively, hereafter, slightly inaccurately, referred to as the North Sea group) were tested in FSTAT 2.9.3 using permutation tests. FSTAT 2.9.3 was also used to estimate differentiation (F_{ST}) between each pair of samples and overall using Weir & Cockerham's (1984) unbiased estimator θ. Pairwise population differentiation was tested using contingency tests implemented in FSTAT 2.9.3. We used PCAGEN 1.3.1 (available at: www2.unil.ch/popgen/softwares/pcagen.htm) to perform a principal component analysis (PCA) based on allele frequencies of all 11 samples and significance of each principal component (PC) was tested by 10 000 randomisations. The proportions of genetic variation distributed between the Baltic Sea and North Sea groups as well as between temporal samples within locations (GER and BOR) were estimated using a hierarchical analysis of molecular variance (AMOVA) implemented in ARLEQUIN 3.11 (Schneider et al. 2000).

Salinity levels on spawning locations exhibit strong relationships with genetic structure in the Atlantic herring from the same area and are the environmental factors with the strongest explanatory power when analysing relationships between different environmental variables and population structure (Bekkevold et al. 2005). To test for such a relationship in sprat partial Mantel tests were applied on all northern samples (i.e. omitting the Adriatic Sea, ADR) to test the correlation between θ values and either geographic distance (shortest waterway distance) or 'environmental distance' (applying difference in mean surface salinity as a proxy) alone, and controlling for each of the explanatory factors. These analyses were performed in FSTAT 2.9.3 using 10 000 randomisations.

RESULTS

Genetic variation

Overall, scoring of genotypes was consistent between persons, months and reanalyses. However, 38 spurious genotypes (usually inconsistent scoring of genotypes and/or consistently weak amplification of fragments) were omitted from further analyses leaving 931 individuals (78 to 88 per population) of which 94.9% of all genotypes were scored successfully (Appendix 1). The MICRO-CHECKER analyses did not suggest any major scoring problems, albeit 28 of 99 tests (28.3%) suggested minor problems with null alleles. Null allele frequencies (r) estimated according to Chakraborty et al. (1992) were in the range of $r = 0.04$ to 0.17 (average = 0.08) and distributed among 8 of 9 loci and all samples. Considering this wide and non-systematic distribution of potential null alleles and the fact that θ values did not appear to be seriously biased by the occurrence of null alleles (see below), genotype frequencies were not corrected before estimating population differentiation. Of 99 tests, 8 tests, distributed over 4 different loci (*Spsp77C*, *Spsp133*, *Spsp154* and *Spsp170*), and 7 samples showed deviations from HWE ($\alpha = 0.05$; Appendix 1) after adjusting for multiple sequential tests (Rice 1989). No significant gametic phase disequilibrium was found across loci and samples after adjusting for multiple sequential tests. A_r varied across loci (Appendix 1). Averaged over-loci estimates of A_r did not vary significantly among samples within the 2 major groups (see Table 1). However, comparing groups of samples within each area (omitting BEL representing the central North Sea–Baltic Sea transition zone) revealed significantly lower genetic diversity in the Baltic Sea samples ($A_r = 14.05 \pm 0.19$ [mean ± SD]) compared with the North Sea group samples ($A_r = 17.10 \pm 0.84$, $p < 0.01$).

Temporal genetic differentiation

No differentiation was found between temporal (2004 and 2005) comparisons from the German Bight ($\theta = 0.002$; 95% confidence interval [CI] = −0.001 to 0.005, $p > 0.05$), while samples from the Bornholm Basin (2005 and 2006) exhibited statistically significant, although low, differentiation ($\theta = 0.006$; 95% CI = 0.002 to 0.010, $p < 0.05$). However, differentiation was statistically non-significant when one or more of the loci exhibiting deviations from HWE were removed (not shown).

Spatial analyses of population differentiation

Due to the minor, but significant, temporal differentiation in the Bornholm Basin and the fact that sample

sizes were large and fairly equal among collections, only the most recent temporal replicates (BOR06 and GER05) were used for spatial comparisons to reduce the overall temporal separation in spatial comparisons. Pairwise θ values for all 9 locations are shown in Table 2 and ranged from 0.001 to 0.089 with an overall θ of 0.030 (95% CI = 0.015 to 0.048, $p < 0.001$). The Adriatic Sea population was highly differentiated from all northern samples (pairwise θ = 0.073 to 0.089, $p < 0.001$ for all comparisons).

Pairwise comparisons between samples within the North Sea group and the Baltic Sea group, respectively, revealed low θ values (between 0.001 and 0.009). Nonetheless, all 3 pairwise tests within the North Sea group and 3 of 6 tests within the Baltic Sea group were statistically significant even after correcting for multiple tests (Table 2). Pairwise comparisons between North Sea and Baltic Sea samples ranged from 0.019 to 0.031, and all were highly significant ($p < 0.001$). The Belt Sea sample was significantly differentiated from all neighbouring samples ($p < 0.001$) and showed a general pattern of intermediate levels of differentiation compared with North Sea–Baltic Sea comparisons (θ = 0.012 to 0.022).This pattern of a strong genetic differentiation between the Baltic Sea and the North Sea mirrored the steep gradient in average surface salinity (Fig. 2).

A potential bias in population differentiation estimates due to 4 loci not exhibiting HWE in several samples (see above) was tested further by recalculating overall, as well as pairwise, θ after omitting each of these loci in turn and when omitting all 4 loci simultaneously. None of these estimates returned greatly differing overall values of θ and estimates obtained by omitting either *Spsp77C*, *Spsp133* or *Spsp154*, respectively, resulted in slightly higher overall θ values (θ = 0.032 to 0.033). Thus, including information from those 3 loci is not expected to inflate estimates of differentiations across regions. Similarly, pairwise θ estimates changed little in any of the reanalyses testing the effect of all 4 loci (see above), although comparisons involving samples from the North Sea group gained statistical significance in a few cases. These minor changes do not warrant the exclusion of any of the 4 loci in the present study but illustrate that great caution should be taken when interpreting low (<0.01) but statistically significant F_{ST} estimates due to high

Table 2. Genetic differentiation (pairwise F_{ST}-values) estimated by θ (Weir & Cockerham 1984), 95% confidence intervals (below diagonal) and p-values (above diagonal). Level of significance obtained following sequential Bonferroni correction for multiple tests (k = 36 tests, Rice 1989). See Table 1 for sample ID. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, ns = not significant

Sample ID	GOT	GDA	BOR06	ARK	BEL	KAT	GER05	CEL	ADR
GOT	–	0.51900	0.02291	0.11207	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
GDA	0.002 ns (0.001–0.003)	–	<0.0001	0.00575	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
BOR06	0.002 ns (–0.001–0.005)	0.006*** (0.001–0.010)	–	0.00242	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
ARK	0.001 ns (–0.001–0.004)	0.004* (0.000–0.009)	0.002* (–0.001–0.005)	–	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
BEL	0.016*** (0.008–0.024)	0.012*** (0.006–0.017)	0.013*** (0.005–0.022)	0.014*** (0.009–0.021)	–	<0.0001	<0.0001	<0.0001	<0.0001
KAT	0.022*** (0.009–0.039)	0.021*** (0.008–0.034)	0.019*** (0.007–0.033)	0.022*** (0.010–0.035)	0.010*** (0.001–0.026)	–	0.00011	0.00187	<0.0001
GER05	0.030*** (0.012–0.055)	0.029*** (0.013–0.048)	0.028*** (0.011–0.047)	0.030*** (0.012–0.049)	0.017*** (0.003–0.036)	0.007** (0.002–0.013)	–	0.00022	<0.0001
CEL	0.031*** (0.016–0.052)	0.029*** (0.018–0.043)	0.027*** (0.015–0.040)	0.028*** (0.016–0.042)	0.022*** (0.008–0.040)	0.009* (0.003–0.015)	0.007** (0.000–0.014)	–	<0.0001
ADR	0.073*** (0.022–0.133)	0.077*** (0.029–0.138)	0.076*** (0.027–0.133)	0.089*** (0.026–0.160)	0.087*** (0.023–0.164)	0.084*** (0.023–0.155)	0.076*** (0.023–0.139)	0.086*** (0.026–0.157)	–

interlocus variability (Chapuis & Estoup 2007, Nielsen et al. 2009), especially in high gene flow scenarios.

Only the first 2 principal components of the PCA explained a significant proportion of the total genetic variance (PC1 and 2, $p < 0.001$; PC3 to 10, $p = 1.000$). The first principal component (PC1, explaining 42% of the variance) in the PCA plot (Fig. 3) mainly separated the ADR population from all others, while PC2 (explaining 32% of the variance) separated samples from the North Sea group and Baltic Sea group into 2 major clusters with the Belt Sea sample located in between (Fig. 3). The Baltic Sea cluster did not reveal any obvious spatial pattern while a weak spatial pattern was evident within the North Sea group. The 2 temporal samples from the German Bight clustered together and exhibited statistically significant differentiation from the other samples in the North Sea area (KAT and CEL; Fig. 3, Table 2). Another PCA omitting the ADR population revealed no further spatial pattern among the remaining samples (not shown). The hierarchical AMOVA grouping temporal samples from the German Bight and Bornholm Basin, respectively, revealed a much higher degree of spatial (2.85%, $p > 0.05$) than temporal (0.15%, $p < 0.05$) genetic variance, although only the temporal comparison was significant. The lack of statistical significance for the former estimate was probably an effect of reduced statistical power in the spatial comparison due to fewer degrees of freedom compared with the temporal comparison (df = 1 and 2, respectively). Another AMOVA (omitting BEL) showed that a significant proportion of the observed genetic variation could be explained by differentiation between the North Sea group and Baltic Sea group (2.21%, $p < 0.05$) while differentiation among locations within these groups explained a much lower part of the overall genetic variation (0.32%, $p < 0.001$). Again, the lower level of statistical significance for the between-group comparison is probably explained by lower statistical power compared with the within-group comparison (df = 1 and 5, respectively). The partial Mantel tests revealed a higher correlation between genetic and environmental (salinity) distance ($r = 0.98$, $p = 0.0001$) than between genetic and geographic distance ($r = 0.63$, $p = 0.0003$). When controlling for environmental distance, the geographic distance parameter became non-significant ($r = 0.63$, $p = 0.71$) while the environmental parameter remained highly significant ($r = 0.76$, $p = 0.0001$) when controlling for geographic distance.

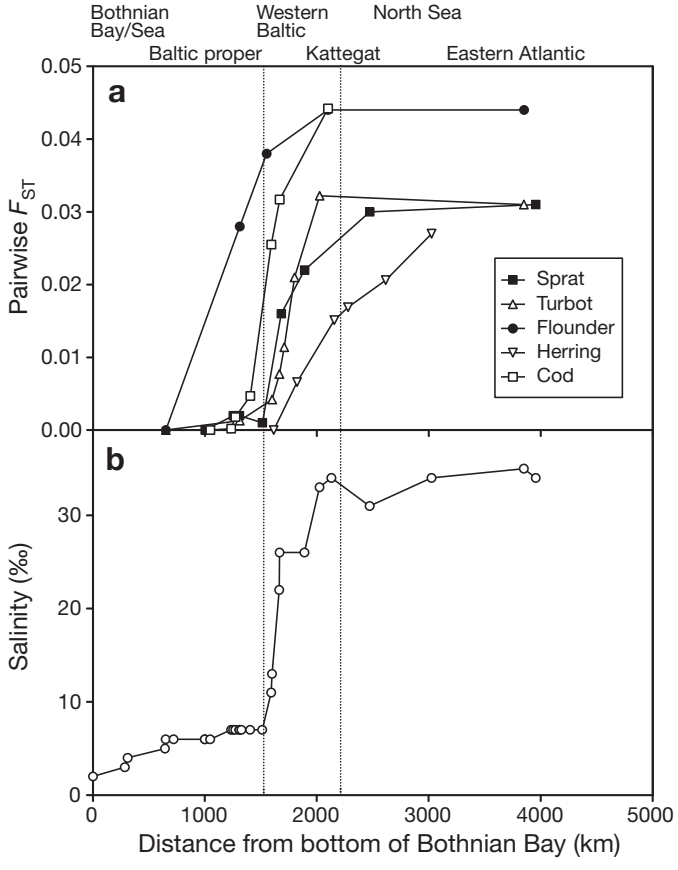


Fig. 2. *Sprattus sprattus*, *Platichthys flesus*, *Gadus morhua*, *Psetta maxima* and *Clupea harengus*. (a) Genetic differentiation (pairwise F_{ST}) between the most northern Baltic sample and samples following a geographical transect from the northern Baltic Sea to the Atlantic Ocean for flounder (Hemmer-Hansen et al. 2007b), cod (Nielsen et al. 2003), turbot (Nielsen et al. 2004), sprat (present study) and herring (Bekkevold et al. 2005). (b) Average surface salinity from the Bothnian Bay to the north-eastern Atlantic Ocean. Vertical lines indicate the area of the transition zone where the salinity gradient is steepest

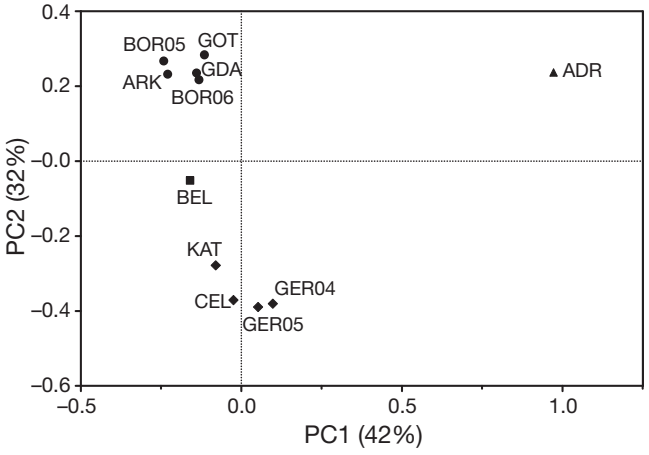


Fig. 3. *Sprattus sprattus*. Plot of first 2 principal components of genetic variation in sprat samples from (●) the Baltic Sea, (■) the Belt Sea, (◆) the North Sea group and (▲) the Adriatic Sea. See Table 1 for sample locations. Principal components (PC) 3 to 10 each explained <6.62% of the total variation. p-values for the proportion of inertia of each axis: PC1 = 0.01, PC2 = 0.01 and PC3 to PC10 = 1.00

DISCUSSION

Large-scale population structure

This is the first study to demonstrate highly significant population structure in European sprat based on highly polymorphic molecular markers. We estimated an overall F_{ST} of 0.030 (95% CI = 0.015 to 0.048), which corresponds with similar scale studies of other marine fishes (Ruzzante et al. 1998, Nielsen et al. 2003, 2004, Bekkevold et al. 2005, Hemmer-Hansen et al. 2007b). We found a sharp genetic division between North Sea and Baltic Sea populations (see below). The Adriatic Sea population exhibited a relatively large divergence from all other samples (Table 2, Fig. 3). This is in concordance with mtDNA data showing evidence for 2 'major clades', with one distributed in the eastern Mediterranean Sea, including the Adriatic Sea and Black Sea, and another in the western Mediterranean Sea, northeast Atlantic Ocean, North Sea and Baltic Sea (Debes et al. 2008). Contrary to the present study, Debes et al. (2008) found no differentiation in allele frequencies among samples ranging from the Bay of Biscay to the Western Baltic. Rather, when grouping samples from the northeast Atlantic, North Sea and Baltic Sea, Debes et al. (2008) found a unimodal mismatch distribution and a 'star-burst'-shaped haplotype network supporting a 'recent' (i.e. following postglacial creation of marine habitats 13 000 to 7600 yr BP) northward range expansion (Debes et al. 2008). These combined results suggest a recent (within the last 10 000 yr) split between northeast Atlantic and Baltic Sea populations. This would translate into approximately 4000 to 5000 sprat generations, which is assumed to be sufficient time for generating the observed levels of genetic differentiation through genetic drift, when considering realistic combinations of effective population size (N_e) and migration rate (m) for marine fishes (Hauser & Carvalho 2008).

Population structure within the North Sea and the Baltic Sea

When studying spatial population structure in high gene flow scenarios, as in most marine fishes, it is vital to define a genetically unique population (Waples & Gaggiotti 2006). Here we apply the weakest criterion from Waples & Gaggiotti (2006) for defining populations from an evolutionary (and not demographic) paradigm, $N_e m < 25$. Choosing a fixed threshold value will always be prone to subjectivity. However, the above threshold conforms to F_{ST} values as low as ~0.01 being statistically highly significant (Waples & Gaggiotti 2006), which is often the case for marine fishes. Thus, despite a few statistically significant pairwise comparisons within the Baltic Sea (Table 2), our data most probably reflect an overall pattern of no spatial structure as inferred from, assumingly, neutral microsatellite markers. No study has revealed evidence of a temporally stable genetic structure of sprat within the Baltic Sea, and a previous approach applying allozyme markers also failed to distinguish among spawning components (Kozlovski 1988).

Within the North Sea/North Atlantic group a weak structure, at most, was detected among populations (Fig. 3, Table 2). These low estimates of spatial differentiation mirror results obtained for other fishes in the area, e.g. herring (Mariani et al. 2005) and flounder (Hemmer-Hansen et al. 2007b). The Celtic Sea sample did, however, not include spawning individuals and our estimate of differentiation might, thus, be an underestimate due to the potential inclusion of transient migrants from other populations. Nonetheless, no study has indicated that such migrations occur. Based on allozyme markers and phenotypic traits, Nævdal (1968) suggested the occurrence of reproductively isolated components of sprat among Norwegian fjords, as also reported in herring (Bekkevold et al. 2005) and cod (Knutsen et al. 2007). However, more detailed sampling is needed for a comprehensive analysis of population structure within the North Sea. The present results should not be interpreted as evidence that sprat in the North Sea and Baltic Sea areas, respectively, are effectively panmictic. For instance, adaptive genetic divergence at genes exposed to local selection can easily be overlooked when studying presumably neutral (i.e. non-functional) variation in a high gene flow scenario (e.g. see Hemmer-Hansen et al. 2007a).

The North Sea–Baltic Sea transition zone

Our results support the existence of a barrier to gene flow separating the northern Kattegat, North Sea and Celtic Sea from Baltic Sea samples (Figs. 2 & 4), with the Belt Sea sample representing a genetically intermediate transition zone. The clustering into 2 regions was further supported by the AMOVA results, which revealed that a higher degree of overall variation was explained by spatial variation between the 2 clusters compared with variation within clusters and between years. If we consider a scenario where 'genetically pure' populations occur in the Baltic Sea and the North Sea area, the narrow transition zone could reflect either a contact zone constituted of genetically admixed individuals (hybrids), or a zone where individuals from the 2 populations mix mechanically (i.e. a Wahlund effect). In theory, the latter scenario will lead to deviations from HWE causing higher than expected inbreed-

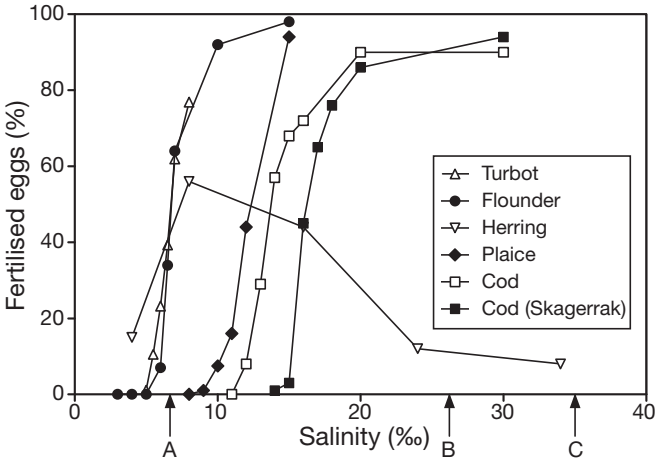


Fig. 4. *Psetta maxima*, *Gadus morhua*, *Platichthys flesus*, *Pleuronectes platessa* and *Clupea harengus*. Fertilisation success as a function of salinity for turbot (Nissling et al. 2006), cod (Nissling & Westin 1997), flounder (with pelagic eggs), plaice *Pleuronectes platessa* (Nissling et al. 2002) and herring (Griffin et al. 1998). All individuals originated from Baltic Sea populations unless otherwise stated. Arrows A, B and C show approximate surface salinities in the Baltic Sea proper, Kattegat and North Sea, respectively

ing coefficients (F_{IS}), but the multi-locus F_{IS} estimate for the BEL sample is not fundamentally different from other samples (Appendix 1). The genetic composition of the BEL sample was further evaluated by simulating a mechanically mixed (50:50) sample of sprat from the 'pure' North Sea and Baltic Sea populations (using the procedure described in Nielsen et al. 2001). This simulated sample did not deviate significantly from HWE at any of the 9 microsatellite loci (results not shown), demonstrating that genetic resolution is too low for statistical detection of a potential Wahlund effect caused by mechanical mixing. Furthermore, calculating individual admixture proportions in the BEL sample, again assuming Baltic and North sea sprat as 'pure' contributing populations (following the procedure described in Nielsen et al. 2003) revealed that the distribution of BEL genotypes resembled a scenario of genetically admixed individuals, rather than mechanical mixing (results not shown).

The partial Mantel tests revealed a stronger correlation between genetic distance and difference in salinity on spawning sites, than between genetic and geographic distance. Results thus clearly demonstrate that isolation by distance is unlikely to account for the observed population structure per se (Figs. 2 & 3). Although we cannot infer causal evolutionary processes directly from simple correlation analyses, our results suggest that salinity differences (and/or some correlated factor) play a role in maintaining reproductive barriers between the North Sea and Baltic Sea. Other factors also probably reinforce reproductive isolation by limiting temporal and/or spatial overlaps of different groups of spawners and juveniles. Such potential factors include physical forcing (Hinrichsen et al. 2005) and environmentally induced spawning behaviour and/or survival (Köster et al. 2003a). However, these possible effects need not be mutually exclusive with an environmentally (e.g. salinity) induced barrier to gene flow.

Reduced allelic richness was observed in Baltic Sea sprat compared with samples from the northern Kattegat, North Sea and Celtic Sea (Table 1). Similar results have been reported for cod (Nielsen et al. 2003), herring (Bekkevold et al. 2005) and flounder (Hemmer-Hansen et al. 2007b). Founding of new populations is commonly associated with loss of genetic variation (Nei et al. 1975). The shallow phylogeography of northeast Atlantic sprat populations (Debes et al. 2008) and the geographically marginal population in the Baltic Sea suggest a recent colonisation by sprat, potentially associated with reinforcement of adaptive divergence in response to low and varying salinity in the Baltic Sea (i.e. a primary contact zone scenario, Garant et al. 2007). However, alternative scenarios, such as a secondary contact zone created during the last glacial retreat (Hewitt 2004, Knowles & Richards 2005), as often suggested for invertebrates in this region (e.g. Väinölä 2003), cannot be ruled out based on the present data.

Interspecific comparison of genetic structure

Our study contributes to existing knowledge of marked genetic clines in marine fishes in the Baltic Sea and adjacent northeastern Atlantic regions. The threshold values in salinity for successful spawning (Fig. 4) further indicate that Baltic Sea components of most marine fishes indeed represent marginal populations with distributional boundaries governed by one or more environmental factors. Cod, in particular, is restricted to the deeper more saline water for successful spawning in the Baltic Sea basins (Fig. 4). Interspecific differences in population structure patterns may reflect variation in, for example, spawning strategy, salinity tolerance (Fig. 4) and/or other traits. In this respect, it is intriguing that the geographic location of the most pronounced barriers to gene flow between Baltic Sea and North Sea populations indeed seems to differ among species (Fig. 2). Bekkevold et al. (2005) showed that on a macro-geographical scale herring exhibits highly significant population structure with differentiation occurring across multiple barriers within the transition zone. This is similar to the 1-dimensional patterns observed for turbot (Nielsen et al. 2004) and

sprat (present study). In these species, genetic divisions mirror surface salinity gradients (Fig. 2). In contrast, areas of most restricted gene flow do not directly correlate with the surface salinity gradient for cod and flounder (Fig. 2). For cod populations, a major division between components of the Western Baltic and the Baltic Sea proper occurs around the Bornholm Basin (Fig. 2, Nielsen et al. 2003). Indeed, the salinity and dissolved oxygen conditions of the Bornholm Basin presently make it the only major area suitable for cod spawning in the Baltic Sea (MacKenzie et al. 2000, Köster et al. 2005). In flounder, neutral genetic differentiation is comparatively low among populations from the Skagerrak and southwestern Baltic, and instead a sharp division is observed near Gotland in the central Baltic Sea (Fig. 2, Hemmer-Hansen et al. 2007b). This division has been ascribed to the occurrence of a shift in life history strategy with populations north of this division having demersal eggs as opposed to pelagic eggs (Nissling et al. 2002). When considering these interspecific differences, one cannot rule out small-scale sampling effects due to different sampling locations among species. However, even with a cautious interpretation we see strong evidence for interesting differences in the pattern of genetic structure, inferred from neutral microsatellites, among the species compared in Fig. 2. This suggests that multi-species approaches in future studies might be rewarding in terms of untangling key evolutionary mechanisms shaping population structure in the sea and e.g. for implementing multispecies stock management. Furthermore, recent studies have provided more direct evidence for the existence of adaptive evolution in the marine environment, despite a background of high gene flow (Hemmer-Hansen et al. 2007a, Larsen et al. 2007, 2008). Inferring the relative importance of external evolutionary drivers and species-specific traits like population history, life history strategy, and migratory and reproductive behaviours remains a great challenge. Therefore, it must be stressed that it is difficult to assess the relative importance of salinity compared with other such factors. Nevertheless, salinity seems to be a key external factor potentially driving evolution and shaping dispersal and population distribution patterns of marine organisms inhabiting the Baltic Sea, including European sprat.

Acknowledgements. The authors are grateful to G. Kraus, R. Hanel, J. P. Hermann, W. Grygiel, G. Kornilovs, H. Haslob, K. Jensen and J. Ellis for help with sample collection, and to N. D. Jong, K. L. D. Mensberg, D. Meldrup and T. B. Christensen for excellent assistance in the laboratory. E. E. Nielsen and 4 anonymous referees are thanked for constructive comments on earlier drafts of this paper. This work is part of the research project UNCOVER (www.uncover.eu, FP6-2004-18 SSP-4, Proposal no. 22717) funded by the European Union within the 6th framework programme. R. Hanel, C. Tsigenopoulos and C. Andre contributed invaluable help and discussion via the European Network of Excellence MARBEF NoE (CT-2003-505446).

LITERATURE CITED

Alheit J (1988) Reproductive biology of sprat (*Sprattus sprattus*): factors determining annual egg production. J Cons Int Explor Mer 44:162–168

Aro E (1989) A review of fish migration in the Baltic. Rapp P-V Réun Cons Int Explor Mer 190:72–96

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. Annu Rev Ecol Syst 16:113–148

Bekkevold D, Andre C, Dahlgren TG, Clausen LAW and others (2005) Environmental correlates of population differentiation in Atlantic herring. Evolution 59:2656–2668

Bonin A, Bellemain E, Eidesen PB, Pompanon F, Brochmann C, Taberlet P (2004) How to track and assess genotyping errors in population genetics studies. Mol Ecol 13: 3261–3273

Chakraborty R, Deandrade M, Daiger SP, Budowle B (1992) Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. Ann Hum Genet 56:45–57

Chapuis MP, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. Mol Biol Evol 24: 621–631

Dailianis T, Limborg M, Hanel R, Bekkevold D, Lagnel J, Magoulas A, Tsigenopoulos CS (2008) Characterization of nine polymorphic microsatellite markers in sprat (*Sprattus sprattus* L.). Mol Ecol Resour 8:861–863

Debes PV, Zachos FE, Hanel R (2008) Mitochondrial phylogeography of the European sprat (*Sprattus sprattus* L., Clupeidae) reveals isolated climatically vulnerable populations in the Mediterranean Sea and range expansion in the northeast Atlantic. Mol Ecol 17:3873–3888

De Silva SS (1973) Aspects of reproductive biology of sprat, *Sprattus sprattus* (L.) in inshore waters of west coast of Scotland. J Fish Biol 5:689–705

Garant D, Forde SE, Hendry AP (2007) The multifarious effects of dispersal and gene flow on contemporary adaptation. Funct Ecol 21:434–443

Goudet J (1995) FSTAT (version 1.2): a computer program to calculate F-statistics. J Hered 86:485–486

Griffin FJ, Pillai MC, Vines CA, Kaaria J, Hibbard-Robbins T, Yanagimachi R, Cherr GN (1998) Effects of salinity on sperm motility, fertilization, and development in the Pacific herring, *Clupea pallasi.* Biol Bull (Woods Hole) 194:25–35

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361–372

Hauser L, Carvalho GR (2008) Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. Fish Fish 9:333–362

Hemmer-Hansen J, Nielsen EE, Frydenberg J, Loeschcke V (2007a) Adaptive divergence in a high gene flow environment: *Hsc70* variation in the European flounder (*Platichthys flesus* L.). Heredity 99:592–600

Hemmer-Hansen J, Nielsen EE, Grønkjær P, Loeschcke V (2007b) Evolutionary mechanisms shaping the genetic population structure of marine fishes; lessons from the European flounder (*Platichthys flesus* L.). Mol Ecol 16: 3104–3118

Hewitt GM (2004) Genetic consequences of climatic oscilla-

tions in the Quaternary. Philos Trans R Soc Lond B Biol Sci 359:183–195

Hinrichsen HH, Kraus G, Voss R, Stepputtis D, Baumann H (2005) The general distribution pattern and mixing probability of Baltic sprat juvenile populations. J Mar Syst 58: 52–66

ICES (2007) Report of the herring assessment working group for the area south of 62°N (HAWG). International Council for the Exploration of the Sea, Copenhagen

Iles TD, Sinclair M (1982) Atlantic herring: stock discreteness and abundance. Science 215:627–633

Johannesson K, Andre C (2006) Life on the margin: genetic isolation and diversity loss in a peripheral marine ecosystem, the Baltic Sea. Mol Ecol 15:2013–2029

Jørgensen HBH, Hansen MM, Bekkevold D, Ruzzante DE, Loeschcke V (2005) Marine landscapes and population genetic structure of herring (*Clupea harengus* L.) in the Baltic Sea. Mol Ecol 14:3219–3234

Knowles LL, Richards CL (2005) Importance of genetic drift during Pleistocene divergence as revealed by analyses of genomic variation. Mol Ecol 14:4023–4032

Knutsen H, Olsen EM, Ciannelli L, Espeland SH and others (2007) Egg distribution, bottom topography and small-scale cod population structure in a coastal marine system. Mar Ecol Prog Ser 333:249–255

Köster FW, Hinrichsen HH, Schnack D, St John MA and others (2003a) Recruitment of Baltic cod and sprat stocks: identification of critical life stages and incorporation of environmental variability into stock-recruitment relationships. Sci Mar 67:129–154

Köster FW, Möllmann C, Neuenfeldt S, Vinther M and others (2003b) Fish stock development in the Central Baltic Sea (1974–1999) in relation to variability in the environment. ICES Mar Sci Symp 219:294–306

Köster F, Möllmann C, Hinrichsen HH, Wieland K and others (2005) Baltic cod recruitment – the impact of climate variability on key processes. ICES J Mar Sci 62:1408–1425

Kozlovski AY (1988) Preliminary results of the studies of the Baltic sprat stock structure using genetic-biochemical methods. Fisch-Forsch 26:74–76 (in Russian)

Larsen PF, Nielsen EE, Williams TD, Hemmer-Hansen J and others (2007) Adaptive differences in gene expression in European flounder (*Platichthys flesus*). Mol Ecol 16: 4674–4683

Larsen PF, Nielsen EE, Williams TD, Loeschcke V (2008) Intraspecific variation in expression of candidate genes for osmoregulation, heme biosynthesis and stress resistance suggests local adaptation in European flounder (*Platichthys flesus*). Heredity 101:247–259

MacKenzie BR, Hinrichsen HH, Plikshs M, Wieland K, Zezera AS (2000) Quantifying environmental heterogeneity: habitat size necessary for successful development of cod *Gadus morhua* eggs in the Baltic Sea. Mar Ecol Prog Ser 193:143–156

Mariani S, Hutchinson WF, Hatfield EMC, Ruzzante DE and others (2005) North Sea herring population structure revealed by microsatellite analysis. Mar Ecol Prog Ser 303: 245–257

Nævdal G (1968) Studies on hemoglobins and serum proteins in sprat from Norwegian waters. Fiskeridir Skr Ser Havunders 14:160–182

Nei M, Maruyama T, Chakraborty R (1975) Bottleneck effect and genetic variability in populations. Evolution 29:1–10

Nielsen EE, Hansen MM, Bach LA (2001) Looking for a needle in a haystack: discovery of indigenous Atlantic salmon (*Salmo salar* L.) in stocked populations. Conserv Gen 2: 219–232

Nielsen EE, Hansen MM, Ruzzante DE, Meldrup D, Grønkjær P (2003) Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea revealed by individual admixture analysis. Mol Ecol 12:1497–1508

Nielsen EE, Nielsen PH, Meldrup D, Hansen MM (2004) Genetic population structure of turbot (*Scophthalmus maximus* L.) supports the presence of multiple hybrid zones for marine fishes in the transition zone between the Baltic Sea and the North Sea. Mol Ecol 13:585–595

Nielsen EE, Wright PJ, Hemmer-Hansen J, Poulsen AN, Gibb IM, Meldrup D (2009) Microgeographical population structure of cod *Gadus morhua* in the North Sea and west of Scotland: the role of sampling loci and individuals. Mar Ecol Prog Ser 376:213–225

Nissling A (2004) Effects of temperature on egg and larval survival of cod (*Gadus morhua*) and sprat (*Sprattus sprattus*) in the Baltic Sea – implications for stock development. Hydrobiologia 514:115–123

Nissling A, Westin L (1997) Salinity requirements for successful spawning of Baltic and Belt Sea cod and the potential for cod stock interactions in the Baltic Sea. Mar Ecol Prog Ser 152:261–271

Nissling A, Westin L, Hjerne O (2002) Reproductive success in relation to salinity for three flatfish species, dab (*Limanda limanda*), plaice (*Pleuronectes platessa*), and flounder (*Pleuronectes flesus*), in the brackish water Baltic Sea. ICES J Mar Sci 59:93–108

Nissling A, Johansson U, Jacobsson M (2006) Effects of salinity and temperature conditions on the reproductive success of turbot (*Scophthalmus maximus*) in the Baltic Sea. Fish Res 80:230–238

Ojaveer E (1989) Population structure of pelagic fishes in the Baltic. Rapp P-V Réun Cons Int Explor Mer 190:17–21

Ojaveer E, Kalejs M (2005) The impact of climate change on the adaptation of marine fish in the Baltic Sea. ICES J Mar Sci 62:1492–1500

Pampoulie C, Gysels ES, Maes GE, Hellemans B, Leentjes V, Jones AG, Volckaert FAM (2004) Evidence for fine-scale genetic structure and estuarine colonisation in a potential high gene flow marine goby (*Pomatoschistus minutus*). Heredity 92:434–445

Parmanne R, Rechlin O, Sjöstrand B (1994) Status and future of herring and sprat stocks in the Baltic Sea. Dana 10: 29–59

Patarnello T, Volckaert FAMJ, Castilho R (2007) Pillars of Hercules: Is the Atlantic–Mediterranean transition a phylogeographical break? Mol Ecol 16:4426–4444

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Hered 86:248–249

Rice WR (1989) Analyzing tables of statistical tests. Evolution 43:223–225

Ruzzante DE, Taggart CT, Cook D (1998) A nuclear DNA basis for shelf- and bank-scale population structure in northwest Atlantic cod (*Gadus morhua*): Labrador to Georges Bank. Mol Ecol 7:1663–1680

Schneider S, Kueffer JM, Roessli D, Excoffier L (2000) ARLEQUIN version 2000: a software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva

Stepputtis D (2006) Distribution patterns of Baltic sprat (*Sprattus sprattus* L.) — causes and consequences. PhD dissertation, Christian Albrechts University, Kiel

Tomkiewicz J, Lehmann KM, St John MA (1998) Oceanographic influences on the distribution of Baltic cod, *Gadus morhua*, during spawning in the Bornholm Basin of the Baltic Sea. Fish Oceanogr 7:48–62

Väinölä R (2003) Repeated trans-Arctic invasions in littoral bivalves: molecular zoogeography of the *Macoma balthica* complex. Mar Biol 143:935–946

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes 4:535–538

Voipio A (1981) The Baltic Sea, Vol 30. Elsevier, Amsterdam

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol Ecol 15:1419–1439

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38: 1358–1370

Whitehead PJP (1985) FAO species catalogue, Vol 7, Part 1. Clupeoid fishes of the world. An annotated and illustrated catalogue of the herrings, sardines, pilchards, sprats, shads, anchovies and wolf-herrings. FAO Fish Synop 125, Rome

Appendix 1. *Sprattus sprattus*. Summary of genetic data for the 11 samples and 9 microsatellite loci analysed. A and A_r are number of alleles and allelic richness (adjusted to n = 56), respectively. H_E and H_O are expected and observed heterozygosity, respectively. Results of tests for deviation from Hardy-Weinberg (HW) proportions are shown by p-values, and significant deviations ($\alpha = 0.05$) after adjustment by the sequential Bonferroni method (Rice 1989) are indicated by asterisks (*). F_{IS} is the estimated inbreeding coefficient (multi-locus estimate given in parentheses). See Table 1 for sample locations

Samples	*Spsp47D*	*Spsp77C*	*Spsp133*	*Spsp154*	*Spsp170*	*Spsp202*	*Spsp219*	*Spsp256*	*Spsp275*
GOT (n = 88)									
% scored	96.6	98.9	98.9	92.0	100	100	95.5	98.9	98.9
A	15	19	10	12	13	10	34	12	15
A_r	14.7	17.4	9.4	11.6	10.6	8.7	30.9	11.2	13.8
H_E	0.820	0.883	0.724	0.851	0.754	0.669	0.945	0.831	0.893
H_O	0.800	0.793	0.644	0.753	0.739	0.761	0.952	0.874	0.874
HW	0.4592	0.0681	0.2218	0.0052	0.2086	0.2731	0.5806	0.019	0.6214
F_{IS} (0.025)	0.025	0.103	0.111	0.115	0.020	–0.139	–0.008	–0.051	0.022
GDA (n = 86)									
% scored	95.3	98.8	97.7	82.6	95.3	98.8	90.7	97.7	98.8
A	17	18	10	11	13	10	29	11	17
A_r	16.0	17.2	9.5	11.0	11.3	8.9	27.1	10.1	15.7
H_E	0.784	0.860	0.739	0.833	0.794	0.708	0.922	0.836	0.888
H_O	0.634	0.847	0.643	0.732	0.817	0.776	0.897	0.833	0.812
HW	0.0024	0.1381	0.0041	0.0001*	0.7845	0.3735	0.2299	0.3863	0.2974
F_{IS} (0.051)	0.192	0.016	0.131	0.121	–0.029	–0.097	0.027	0.003	0.086
BOR05 (n = 82)									
% scored	92.7	85.4	96.3	82.9	95.1	100	89.0	100	96.3
A	16	17	10	13	13	11	30	11	18
A_r	15.2	16.4	9.3	12.4	11.0	9.6	27.9	10.2	17.0
H_E	0.744	0.851	0.752	0.816	0.756	0.667	0.939	0.816	0.884
H_O	0.724	0.786	0.671	0.647	0.756	0.720	0.932	0.780	0.810
HW	0.1834	0.0014	0.0690	0.0009	0.1686	0.2133	0.3060	0.7501	0.1391
F_{IS} (0.056)	0.028	0.077	0.109	0.208	0.000	–0.079	0.008	0.043	0.084
BOR06 (n = 88)									
% scored	98.9	96.6	100	92.0	100	100	95.5	97.7	97.7
A	17	20	7	12	14	10	28	13	15
A_r	16.0	19.0	6.6	11.3	11.5	8.9	25.2	12.1	13.6
H_E	0.809	0.911	0.713	0.848	0.801	0.692	0.927	0.863	0.877
H_O	0.678	0.788	0.568	0.741	0.818	0.682	0.952	0.872	0.860
HW	0.0059	0.0076	0.0004*	0.0034	0.5786	0.2434	0.6609	0.0936	0.6224
F_{IS} (0.065)	0.162	0.135	0.204	0.127	–0.022	0.014	–0.028	–0.010	0.019

Appendix 1 (continued)

Samples	*Spsp47D*	*Spsp77C*	*Spsp133*	*Spsp154*	*Spsp170*	*Spsp202*	*Spsp219*	*Spsp256*	*Spsp275*
ARK (n = 78)									
% scored	98.7	97.4	94.9	76.9	94.9	97.4	84.6	96.2	89.7
A	16	19	10	13	15	7	27	13	14
A_r	15.4	17.5	9.3	12.9	13.4	6.8	25.7	12.1	13.6
H_E	0.806	0.871	0.681	0.869	0.796	0.616	0.931	0.841	0.875
H_O	0.792	0.803	0.649	0.717	0.757	0.645	0.909	0.920	0.829
HW	0.0881	0.2480	0.2621	0.0337	0.6923	0.7540	0.4396	0.0162	0.1073
F_{IS} (0.037)	0.018	0.079	0.048	0.177	0.050	–0.047	0.024	–0.094	0.053
BEL (n = 83)									
% scored	88.0	89.2	92.8	90.4	88.0	94.0	88.0	98.8	96.4
A	13	22	12	11	21	10	30	15	20
A_r	13.0	20.8	11.3	10.9	19.8	9.0	28.3	14.2	18.5
H_E	0.843	0.919	0.755	0.790	0.901	0.652	0.955	0.860	0.841
H_O	0.699	0.877	0.584	0.699	0.767	0.680	0.986	0.805	0.825
HW	0.0102	0.0013	0.0005*	0.0287	0.0218	0.5818	0.6733	0.6285	0.0161
F_{IS} (0.079)	0.171	0.046	0.226	0.116	0.148	–0.042	–0.032	0.065	0.020
KAT (n = 81)									
% scored	90.1	92.6	95.1	69.1	87.7	96.3	91.4	97.5	96.3
A	19	22	10	11	22	14	28	13	22
A_r	17.7	20.1	9.1	11.0	20.9	12.1	26.1	12.3	20.7
H_E	0.849	0.903	0.717	0.817	0.906	0.717	0.935	0.834	0.898
H_O	0.822	0.840	0.636	0.804	0.662	0.808	0.946	0.797	0.846
HW	0.1578	0.1158	0.1962	0.1403	0.0000*	0.0290	0.0180	0.6864	0.0877
F_{IS} (0.055)	0.032	0.070	0.113	0.016	0.271	–0.128	–0.012	0.045	0.058
GER04 (n = 88)									
% scored	100	89.8	98.9	97.7	100	98.9	100	100	100
A	18	24	12	12	28	11	37	15	25
A_r	16.7	21.9	11.0	11.7	26.2	10.1	32.5	13.4	22.7
H_E	0.879	0.923	0.778	0.782	0.952	0.790	0.957	0.848	0.907
H_O	0.852	0.785	0.644	0.674	0.898	0.759	0.920	0.807	0.875
HW	0.4506	0.0005*	0.0199	0.0003*	0.1982	0.1296	0.1254	0.046	0.6384
F_{IS} (0.077)	0.031	0.150	0.173	0.138	0.057	0.040	0.039	0.049	0.035
GER05 (n = 87)									
% scored	95.4	96.6	97.7	97.7	86.2	97.7	98.9	100	97.7
A	19	19	13	11	28	12	35	17	23
A_r	17.4	18.5	11.2	10.4	26.2	10.9	30.9	15.4	21.7
H_E	0.823	0.927	0.717	0.725	0.945	0.780	0.953	0.854	0.921
H_O	0.807	0.857	0.565	0.718	0.907	0.718	0.942	0.816	0.859
HW	0.0613	0.3379	0.0011	0.0886	0.3200	0.0873	0.0680	0.1800	0.0166
F_{IS} (0.060)	0.020	0.076	0.213	0.010	0.040	0.080	0.012	0.045	0.068
CEL (n = 85)									
% scored	92.9	89.4	96.5	83.5	90.6	92.9	91.8	94.1	95.3
A	19	19	10	9	26	11	30	17	22
A_r	17.2	18.0	8.6	8.8	23.1	10.6	26.9	15.6	20.3
H_E	0.836	0.902	0.636	0.736	0.915	0.740	0.942	0.826	0.902
H_O	0.785	0.842	0.634	0.690	0.688	0.772	0.859	0.825	0.877
HW	0.0179	0.0104	0.3654	0.0331	0.0000*	0.0038	0.0105	0.3446	0.6927
F_{IS} (0.062)	0.062	0.066	0.002	0.063	0.249	–0.044	0.088	0.001	0.028
ADR (n = 85)									
% scored	98.8	98.8	100	84.7	100	97.6	100	100	100
A	14	26	10	11	22	20	33	15	10
A_r	13.0	24.1	9.2	10.6	19.4	18.1	28.7	14.1	8.7
H_E	0.724	0.942	0.724	0.844	0.749	0.882	0.943	0.803	0.455
H_O	0.690	0.929	0.506	0.833	0.706	0.735	0.894	0.729	0.388
HW	0.0021	0.1177	0.0002*	0.4476	0.0734	0.0019	0.3465	0.4068	0.0070
F_{IS} (0.093)	0.047	0.015	0.303	0.013	0.058	0.168	0.052	0.092	0.147

Editorial responsibility: Philippe Borsa, Nouméa, France

Submitted: August 27, 2008; Accepted: December 11, 2008
Proofs received from author(s): March 9, 2009

Chapter 4

Microsatellite DNA reveals population genetic differentiation among sprat (*Sprattus sprattus*) sampled throughout the Northeast Atlantic, including Norwegian fjords

Published in *ICES Journal of Marine Science*

ICES Journal of Marine Science (2011), 68(10), 2145–2151. doi:10.1093/icesjms/fsr153

Microsatellite DNA reveals population genetic differentiation among sprat (*Sprattus sprattus*) sampled throughout the Northeast Atlantic, including Norwegian fjords

Kevin A. Glover[1]*, Øystein Skaala[1], Morten Limborg[2], Cecilie Kvamme[1], and Else Torstensen[1]

[1]*Institute of Marine Research, PO Box 1870 Nordnes, N-5817 Bergen, Norway*
[2]*National Institute of Aquatic Resources, Technical University of Denmark, Vejlsøvej 39, DK-8600 Silkeborg, Denmark*

**Corresponding Author: tel: +47 55 236357; e-mail: Kevin.glover@imr.no*

Glover, K. A., Skaala, Ø., Limborg, M., Kvamme, C., and Torstensen, E. Microsatellite DNA reveals population genetic differentiation among sprat (*Sprattus sprattus*) sampled throughout the Northeast Atlantic, including Norwegian fjords. – ICES Journal of Marine Science, 68: 2145–2151.

Received 6 May 2011; accepted 25 August 2011

Sprat (*Sprattus sprattus*), small pelagic shoaling fish, were sampled from the Celtic, North, and Baltic seas, and 10 Norwegian fjords. Significant overall genetic differentiation was observed among samples when analysed with eight microsatellite DNA loci (Global $F_{ST} = 0.0065$, $p < 0.0001$). The greatest genetic differences were observed between the Baltic and all other samples (largest pairwise $F_{ST} = 0.043$, $p < 0.0001$). No significant genetic differentiation was observed between a sample from the Celtic Sea (CEL) and the North Sea (NSEA; $F_{ST} = 0.001$, $p = 0.16$), but variable levels of genetic differentiation were observed among samples collected from Norwegian fjords (pairwise F_{ST} ranging from 0 to 0.0096, most non-significant). All fjord samples were significantly differentiated to NSEA and CEL samples. Further, all fjord samples displayed reduced allelic richness compared with NSEA and CEL samples. Clearly, sprat display population genetic differentiation throughout the Northeast Atlantic, and there may be limited connectivity between Norwegian fjord and sea-going populations.

Keywords: fishery genetics, management, North Sea, pelagic fish, population genetics.

Introduction

The European sprat (hereafter referred to as sprat; *Sprattus sprattus*) is a small, oily pelagic shoaling fish inhabiting the Baltic, the Northeast Atlantic down to Morocco, and the northern Mediterranean basins, as well as the Black Sea (see the ICES FishMap for sprat). Northern Norway is the northernmost latitude to which the species is distributed. The sprat has been and continues to remain important in marine fisheries, sustaining catches between 100 000 and 200 000 t from the North Sea (NSEA)–Skagerrak in the period 1996 to date (ICES, 2011). Catch data for sprat before 1996 are considered unreliable because of a large, but unknown, bycatch of juvenile herring (*Clupea harengus*; ICES, 2011). The sprat fishery is largely opportunistic, and is often more influenced by the abundance of other target species than by sprat abundance alone (ICES, 2009). A commercial fishery targets sprat in the Norwegian fjords, yielding annual catches of 8000–16 000 t in the 1960s, and a peak of 18 000 t in 1972. Thereafter, the catches declined steadily until they stabilized at a low level in the 2000s (1400–3500 t annually; Official statistics, Norwegian Directorate of Fisheries).

The translation of genetic data in fisheries management has not been without its challenges (Waples *et al.*, 2008). However, for fully informed fishery management, it is necessary to quantify and understand the level of population genetic structure displayed within a given species. For example, is the commercial harvest conducted on one or more populations, and do those populations overlap in time and space? Marine fish, with their clear potential for long-distance dispersal, were once regarded as having limited population genetic structure, but it is increasingly evident that there is significant population genetic structure. For instance, a population structure has been observed in well-studied groundfish such as Atlantic cod (*Gadus morhua*; Knutsen *et al.*, 2003; Pampoulie *et al.*, 2006). Moreover, a population genetic structure has been observed in such highly mobile pelagic species as mackerel *Scomber* spp. (Zardoya *et al.*, 2004) and herring (Bekkevold *et al.*, 2005; Jørgensen *et al.*, 2005).

The population genetic structure of sprat was first addressed in two pilot studies using haemoglobin and allozyme genetic variation in Norway (Nævdal, 1968; Jørstad and Nævdal, 1981). Among samples of sprat collected in Norwegian fjords, there was some indication of population genetic structuring. However, despite a suggestion that sprat in Norwegian waters possibly consisted of two or more reproductively isolated populations, no geographic trend in genetic pattern was identified (Nævdal, 1968). Looking at the broader range of the species, a mitochondria-based phylogeographic study of sprat from the Baltic Sea to the Black Sea, via the NSEA and the Mediterranean, revealed two major clades, separated across the Strait of Sicily (Debes *et al.*, 2008). Recently, highly polymorphic microsatellite DNA markers have been developed for sprat (Dailianis *et al.*, 2008), and they have been used to investigate the population genetic structure of sprat across a salinity gradient from the inner regions of the Baltic Sea

to the southern NSEA (Limborg *et al.*, 2009). Those authors noted a highly significant population genetic structure across the environmental gradient between the Baltic Sea and the NSEA, but at most a very weak structure within oceanic basins.

Sprat from the most northern region of its distribution, including sprat within the Norwegian fjords, have to date not been subjected to analysis with DNA markers. Given the fact that the region has sustained a locally important fishery for this species, and that some evidence of population genetic structure has been revealed previously in pilot studies conducted in Norwegian fjords (Nævdal, 1968; Jørstad and Nævdal, 1981), it was thought important to investigate samples from the region with the newly developed microsatellite markers. Consequently, the aim of this study was to investigate the population genetic structure of sprat, with emphasis on the genetic relationship between the larger oceanic populations, in the NSEA and Baltic Sea, and the sprat captured in Norwegian fjords.

Material and methods

The study is based on the analysis of 1025 sprat collected from 14 locations (Table 1; Figure 1). Except Nordfjord and Baltic Sea Gotland, all samples were taken in autumn and not during the spawning season (spring to early summer). Sampling during a spawning season represents the most robust approach to delineating population genetic structure, but most of the samples analysed here were collected as part of the Institute of Marine Research's (IMR's) scientific cruise for fjord sprats (see Table 1). In 2008, nine samples (31–100 fish per sample) were collected from seven Norwegian fjords. Data from the IMR cruise in November–December 2008 confirmed that all fjord sprat $\leq$8 cm were 0-year-olds, whereas larger sprat ($\geq$8.5 cm) belonged to the 1+ group. A further two samples from Norwegian fjords, also taken on board IMR's vessels, were collected in 2001 (Nordfjord) and in 2007 (West Fornebu). In addition to 11 fjord samples, samples of sprat were collected from commercial vessels operating in the NSEA (2008) and the Celtic Sea (CEL; 2009; Figure 1). Finally, a subset of sprat constituting the Baltic Sea Gotland (2006) sample from the study of the species in the Baltic Sea (Limborg *et al.*, 2009) was also included in the analyses. Hereafter, source codes for the sprat samples are generally used, as depicted in Table 1.

Genotyping

DNA was extracted in a 96-well format using the Qiagen DNeasy kit. Each 96-well tray contained a minimum of two blank controls. Eight dinucleotide microsatellite loci developed for sprat (Dailianis *et al.*, 2008) were amplified in two multiplex reactions: Multiplex 1—*Spsp047*, *Spsp077*, *Spsp170*, *Spsp202*; Multiplex 2—*Spsp219*, *Spsp133*, *Spsp256*, *Spsp275*. These markers were amplified using a slightly modified version (i.e. optimizing primer concentrations and using different reagent suppliers) of a protocol described previously (Dailianis *et al.*, 2008). PCR fragments were separated on an ABI 3730 sequencer and sized relative to the Applied Biosystem GeneScan^{TM}–500LIZ^{TM} size standard. Alleles were scored using automatic binning implemented in the Genemapper software (v4.0). Allele scoring was controlled independently by two persons.

Statistical analysis

The program MSA (Dieringer and Schlotterer, 2003) was used to compute summary statistics and values of the fixation index (F_{ST}; a measure of genetic distance among populations). Genepop (Raymond and Rousset, 1995) was used to test for deviation from Hardy–Weinberg equilibrium (HWE) for all loci within each sample. HWE is the state at which genotype frequencies in a population remain stable. This was examined statistically by Fisher's exact test (dememorization 10 000; 100 batches; 5000 iterations). The significance level was presented at $\alpha = 0.05$, in addition to applying Bonferroni correction for multiple tests (Rice, 1989). Genepop was also used to estimate observed (H_o)

Table 1. Locations, dates, and biological information relating to the sprat samples

Sample		Geographic position			Date		Sampling method		Biological information			
Code	Area	North	East	Year	Date	Time, UTC	Vessel	Gear	Sample (*n*)	Mean *L* (cm)	Mean *W* (g)	0-group (%)
Norwegian fjord samples												
LYS	Forsand, Lysefjord	58.92	06.09	2008	12 November	23:55	A	A	100	12.4	12.5	0
HAR-1	Tyssedal, Sørfjord	60.14	06.56	2008	15 November	07:47	A	A	100	9.6	5.2	0
HAR-2	Outer Sørfjord	60.41	06.67	2008	15 November	02:39	A	A	48	6.2	1.1	100
NOR	Nordfjord	61.85	05.85	2001	22 May	06:52	B	A	75	13.6	15.9	0
OSL	West Fornebu	59.89	10.59	2007	30 September	–	C	B	99	12.2	13.7	0
SOG	Skjolden, Sognefjord	61.49	07.59	2008	22 November	16:26	A	A*	49	9.3	4.9	14
HOL	Holandsfjord	66.71	13.63	2008	11 December	16:51	A	A	31	11.1	7.9	0
TRH	Stjørdalsfjord	63.47	10.86	2008	03 December	16:16	A	A	80	11.1	9.1	0
HAR-3	Tittelsnes, Bømlafjord	59.74	05.56	2008	14 November	02:14	A	A	80	6.5	1.7	100
FIN	Rana, Finneidfjord	66.21	13.81	2008	13 December	06:11	A	A	81	7.5	2.7	94
MEL	Melfjord	66.61	13.58	2008	12 December	02:48	A	A*	80	7.3	2.2	100
Sea samples												
NSEA	Southwest North Sea	53.75	01.50	2008	–	–	D	–	94	–	–	–
CEL	Southwest Celtic Sea	52.80	10.08	2009	08 October	–	D	–	76	–	–	–
GOT	Baltic Sea, Gotland	58.24	20.31	2006	May	–	D	–	52	–	–	–

Gear: A, Harstad trawl with floats; A*, Harstad trawl without floats; B, beach-seine.
Vessel: A, "Håkon Mosby" (IMR vessel); B, "Michael Sars" (IMR vessel); C, "G. M. Dannevig" (IMR vessel); D, commercial vessel; Mean *L* and Mean *W*, mean length and mean weight of sprat in that sample, 0-group (%), percentage of age-0 fish in the sample (the other component is fish aged 1+). "–", not known or not applicable.

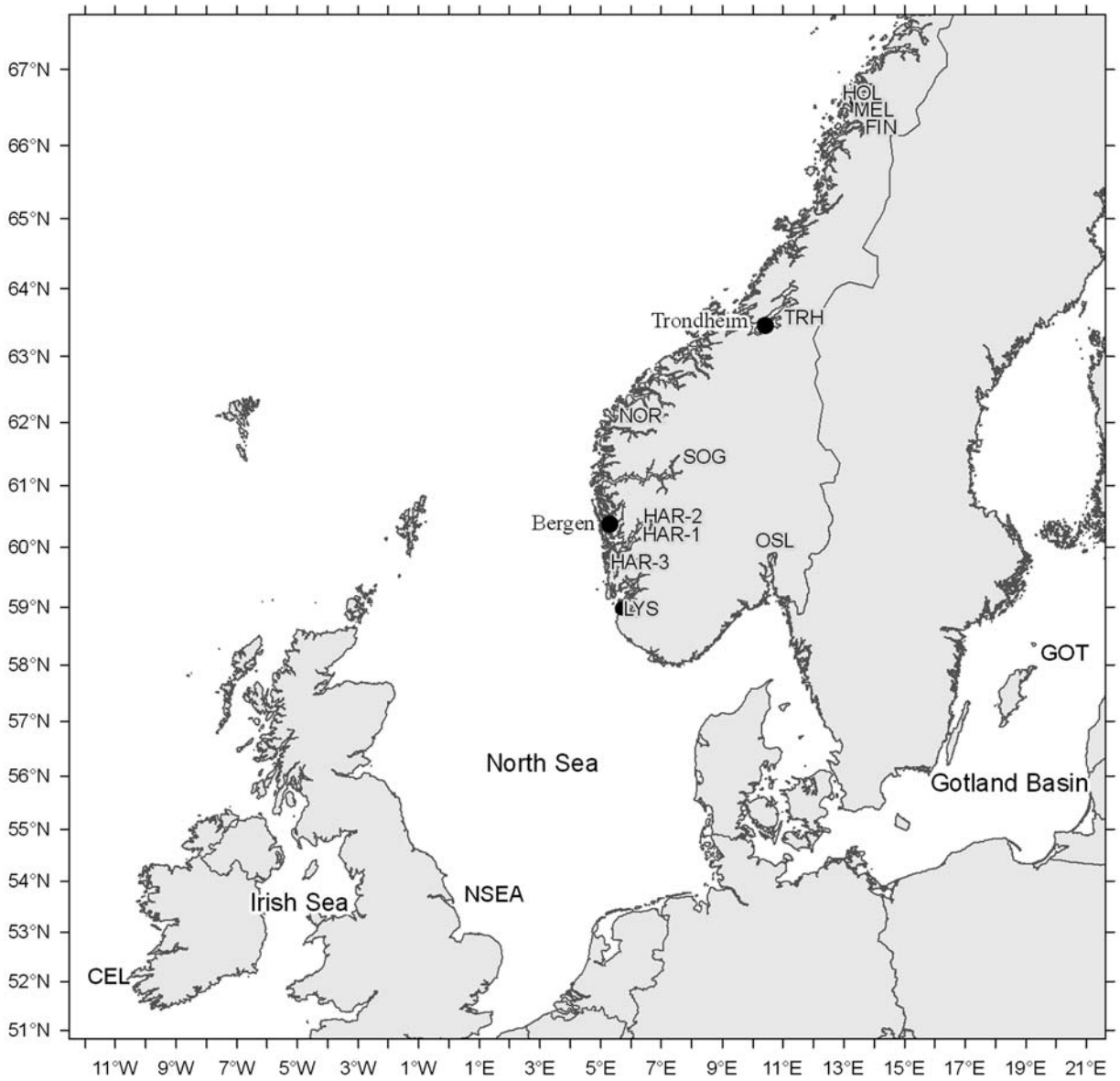


Figure 1. Locations of sprat sample collection used for this study. For further detail, see Table 1.

and expected (H_e) heterozygosities (i.e. the fraction of individuals that are heterozygous in a population) and the inbreeding coefficient F_{IS}.

Estimates of neutral demographic processes can be biased if one or more assumingly neutral genetic markers deviate from neutrality as a result of hitch-hiking selection (Neilsen *et al.*, 2006). Therefore, the program LOSITAN (Antao *et al.*, 2008) was used to test the loci for neutrality. This program utilizes an F_{ST} outlier-detection method to identify loci that are potential candidates for balancing or positive selection. To test for a general pattern of reduced genetic diversity in all the more isolated fjord populations compared with the assumed larger Atlantic populations (CEL and NSEA), we applied the permutation test implemented in FSTAT v2.9.3 (Goudet, 1995) to compare levels of allelic richness (R_S) between these groups. A one-sided test assuming reduced diversity in fjord populations and using 2000 permutations was computed.

To identify any potential cryptic genetic variation within and among samples, Bayesian clustering analysis, as implemented in the program STRUCTURE 2.2 (Pritchard *et al.*, 2000; Falush *et al.*, 2003), was used to assign individual fish to groups without using prior information about their origin. Runs were conducted for the number of putative populations (i.e. k), set at 1–5, each with five iterations. Correlated allele frequencies and an admixture model were assumed. Each run consisted of a burn-in of 250 000 Markov-chain Monte Carlo steps, followed by 1 000 000 steps. MEGA (Tamura *et al.*, 2007) was used to produce phylogenetic trees using the UPGMA (unweighted pair group method with arithmetic mean) method on matrices of pairwise F_{ST} values. The trees were linearized assuming equal evolutionary rates in all lineages (Takezaki *et al.*, 1995).

Results

Dataset and HWE

From in all 8200 genotypes potentially scored (8 loci × 1025 samples), >95% genotyping coverage was achieved. Fish

displaying genetic data for three or fewer loci were excluded from statistical analyses.

From a total of 112 tests (all loci in all samples), 37 significant deviations from HWE were observed at $p = 0.05$ (Table 1). When adjusted for Bonferroni correction (8 loci; critical $p = 0.006$), 21 tests remained significant. These were unevenly distributed, with three loci responsible for most of the deviations: *Spsp077* = 7, *Spsp133* = 8, *Spsp275* = 2 (all other markers yielded one or fewer deviations post-correction). These deviations are most likely caused by null alleles, as reported previously for these markers (Limborg *et al.*, 2009). Consequently, all statistical analyses were conducted on the full set of eight markers in addition to the subset of five markers excluding those displaying extensive deviation from HWE.

Looking specifically at samples, HWE was distributed unevenly among them, with the HOL and FIN samples displaying no deviations at $p = 0.05$, and the HAR-3, OSL, and HAR-1 samples each displaying deviation from HWE in four markers at $p = 0.05$. Following Bonferroni correction and removal of the three loci deviating extensively from HWE, only a single deviation observed in the OSL sample remained.

Within-sample genetic variation

All samples displayed a high degree of allelic variation in all loci, ranging from a low of eight alleles in the Baltic Sea sample for locus *Spsp170* to a high of 42 alleles in the NSEA, OSL, and HAR-1 samples for locus *Spsp219*. The total number of alleles observed for all eight loci pooled ranged from a low of 125 in the Gotland Basin (GOT) sample from the Baltic Sea to 209 in the NSEA sample.

Allelic richness, which partially corrects for the differences in sample size biasing allelic diversity estimates, modified the observed trend slightly, although the GOT (107) and the NSEA (146) samples still represented the samples with least and greatest allelic diversity in the dataset, respectively. Samples taken from the Norwegian fjords were similar in allelic richness to each other. Allelic richness in the fjord populations ($R_S = 16.2$) was significantly reduced compared with the group of two Atlantic samples (i.e. CEL and NSEA; $R_S = 17.5$, $p = 0.012$).

Expected heterozygosity (pooled over all eight loci) displayed little variation among samples, ranging from 0.82 in the Baltic Sea sample to 0.89 in the NSEA and SOG samples. The inbreeding coefficient F_{IS} (pooled over eight loci) displayed positive values in all samples except the Baltic Sea and Holandsfjord samples (both −0.02). The positive F_{IS} values were quite large for some samples, suggesting too few heterozygotes. When computed for the reduced set of five loci (excluding *Spsp077*, *Spsp133*, and *Spsp275*), F_{IS} values decreased markedly and the observed heterozygosities per sample were much closer to the expected ones (Table 2).

Among-sample genetic differentiation

Overall, significant genetic differentiation was observed among the 14 samples (8 loci global $F_{ST} = 0.0065$, $p = 0.0001$; 5 loci global $F_{ST} = 0.0062$, $p = 0.0001$). Loci displayed variable global F_{ST} values: *Spsp*047 = 0.004, $p < 0.0001$; *Spsp077* = 0.0004, $p = 0.27$; *Spsp170* = 0.011, $p < 0.0001$; *Spsp202* = 0.014, $p < 0.0001$; *Spsp133* = 0.004, $p = 0.018$; *Spsp219* = 0.0015, $p = 0.007$; *Spsp256* = 0.0016, $p = 0.087$; *Spsp275* = 0.016, $p < 0.0001$. Following tests of neutrality using the program LOSITAN, the locus *Spsp275* was identified as a potential candidate for positive selection. Of the five loci included in the reduced set of markers conforming to HWE, informative loci still remained.

Pairwise F_{ST} values computed using all eight loci, and the subset of five loci, revealed variable and, in some instances, highly significant differences among samples (Table 3). In both cases, the sample originating in the Baltic Sea displayed the greatest differentiation of any sample (highest eight loci pairwise $F_{ST} = 0.038$, GOT vs. HOL; highest five loci pairwise $F_{ST} = 0.043$, GOT vs. HOL). Although small differences in relationships among samples were observed for the datasets including eight and five loci, the overall pattern of relationships was largely similar (Figure 2). Specifically, the GOT sample from the Baltic Sea was most distinct, the samples from the CEL and the NSEA clustered together, and differences among fjord samples were

Table 2. Within-sample genetic diversity parameters

		Data from eight loci							Data from five loci			
Sample	***n***	**At**	**AtLR**	**Ar**	H_o	H_e	F_{IS}	**HWE**	H_o	H_e	F_{IS}	**HWE**
LYS (1)	100	195	14–39	132	0.80	0.87	0.09	3 (2)	0.86	0.89	0.02	1 (0)
HAR-1 (2)	99	196	13–42	134	0.81	0.87	0.07	4 (1)	0.86	0.88	0.02	1 (0)
HAR-2 (3)	41	134	10–30	129	0.83	0.87	0.04	0 (0)	0.85	0.86	0.02	0 (0)
NOR (4)	74	186	12–37	135	0.81	0.87	0.08	3 (2)	0.89	0.89	0.00	0 (0)
OSL (5)	90	192	14–42	136	0.81	0.87	0.07	4 (3)	0.87	0.88	0.02	1 (1)
SOG (6)	49	150	10–30	130	0.81	0.89	0.09	3 (1)	0.87	0.88	0.01	0 (0)
HOL (7)	31	130	10–25	130	0.88	0.87	−0.02	0 (0)	0.92	0.87	−0.06	0 (0)
TRH (8)	80	193	11–40	139	0.80	0.88	0.09	3 (1)	0.85	0.89	0.05	1 (0)
HAR-3 (9)	80	173	11–34	129	0.78	0.87	0.11	4 (4)	0.83	0.88	0.05	1 (1)
FIN (10)	79	185	12–37	133	0.80	0.88	0.09	4 (2)	0.87	0.89	0.02	1 (0)
MEL (11)	80	190	11–39	138	0.80	0.88	0.09	3 (2)	0.86	0.88	0.03	0 (0)
NSEA (12)	94	209	13–42	146	0.79	0.89	0.11	3 (2)	0.84	0.89	0.05	0 (0)
CEL (13)	76	194	12–40	143	0.83	0.88	0.06	2 (1)	0.88	0.90	0.02	0 (0)
GOT (14)	52	125	8–26	107	0.84	0.82	−0.02	1 (0)	0.86	0.81	−0.06	1 (0)

n, number of fish analysed per sample (note, these values do not necessarily match those in Table 1 because some fish were not used for DNA analysis); At, total number of alleles observed; AtLR, range in total number of alleles per locus; Ar, allelic richness based upon sampling 28–31 fish per locus; H_o, observed heterozygosity; H_e, expected heterozygosity; F_{IS}, inbreeding coefficient; HWE, number of deviations from HWE, with data following Bonferroni correction for multiple tests in parenthesis.

Table 3. Pairwise F_{ST} values among 14 samples of sprat (bottom left matrix), and associated p values (upper right matrix), based on data from eight microsatellite loci

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14
LYS (1)		0.0001	0.014	0.002	0.0001	0.038	0.004	0.0001	0.36	0.0004	0.0001	0.0001	0.0001	0.0001
HAR-1 (2)	0.0054		0.14	0.31	0.012	0.043	0.034	0.54	0.44	0.26	0.28	0.0001	0.0001	0.0001
HAR-2 (3)	0.0046	0.0018		0.022	0.18	0.38	0.44	0.12	0.93	0.1	0.010	0.0001	0.0001	0.0001
NOR (4)	0.0046	0.0005	0.0044		0.002	0.14	0.015	0.20	0.41	0.60	0.32	0.0001	0.0001	0.0001
OSL (5)	0.0096	0.0026	0.0015	0.0044		0.013	0.0006	0.16	0.063	0.017	0.0028	0.0001	0.0001	0.0001
SOG (6)	0.0028	0.0027	0.0005	0.0018	0.0041		0.074	0.13	0.58	0.65	0.23	0.0003	0.0001	0.0001
HOL (7)	0.0063	0.0038	0.0001	0.0053	0.0087	0.0036		0.48	0.39	0.008	0.055	0.0001	0.0001	0.0001
TRH (8)	0.0068	−0.0002	0.0022	0.0009	0.0011	0.0017	−0.0001		0.37	0.55	0.55	0.0001	0.0003	0.0001
HAR-3 (9)	0.0003	0.0001	−0.0026	0.0002	0.0019	−0.0004	0.0004	0.0003		0.51	0.31	0.0001	0.0001	0.0001
FIN (10)	0.0053	0.0006	0.0021	−0.0004	0.0029	−0.0007	0.0056	−0.0003	−0.0001		0.273	0.0002	0.0001	0.0001
MEL (11)	0.0058	0.0005	0.0054	0.0005	0.0040	0.0011	0.0036	−0.0002	0.0005	0.0006		0.0001	0.0001	0.0001
NSEA (12)	0.0147	0.0089	0.0160	0.0080	0.0088	0.0066	0.0142	0.0071	0.0100	0.0058	0.0070		0.16	0.0001
CEL (13)	0.0158	0.0081	0.0184	0.0086	0.0094	0.0100	0.0133	0.0062	0.0101	0.0066	0.0075	0.0011		0.0001
GOT (14)	0.0243	0.0234	0.0219	0.0204	0.0186	0.0254	0.0375	0.0248	0.0221	0.0203	0.0250	0.0279	0.0284	

Underlined values of p remain significant following Bonferroni correction for multiple testing (adjusted critical $p = 0.00055$).

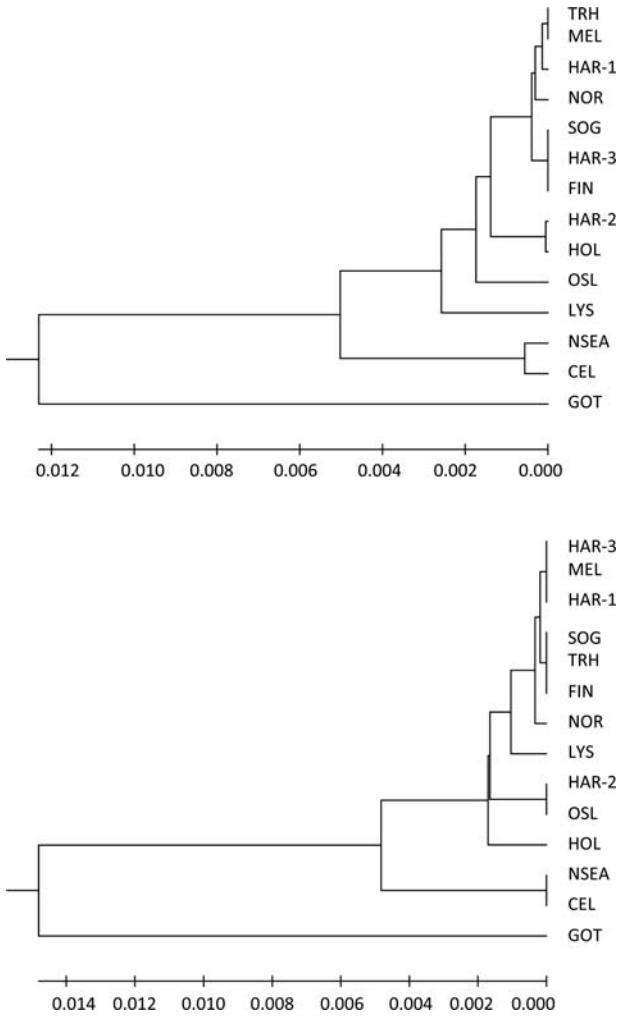


Figure 2. UPGMA diagrams showing genetic relationships among the samples based upon a matrix of F_{ST} values computed using eight (top) and five (bottom) loci.

smaller (although not without exception), largely non-significant (Table 3), and significantly different from the CEL, NSEA, and GOT samples.

Bayesian clustering analysis failed to detect any significant genetic differentiation among samples, or any cryptic genetic structure (data not presented).

Discussion

This is the first DNA-based population genetic analysis comparing sprat sampled in Norwegian fjords, the species' most northern distribution, and the surrounding seas. Although statistically significant population genetic structure was observed throughout the sampling range, only weak evidence of population genetic structure was observed among the samples collected in different Norwegian fjords. The sample from the GOT displayed the largest genetic differences of all samples, with the next largest genetic differences originating in various combinations of fjord samples compared with either the NSEA or CEL samples. The last two samples were not statistically different from each other. Together, these data show that sprat display distinct, and statistically significant, population genetic differentiation among the three major regions sampled here: the Norwegian fjords, the NSEA and CEL, and the Baltic Sea, represented by a single sample from the GOT.

The sample originating in the GOT displayed the lowest genetic diversity, estimated by the total number of alleles pooled over eight loci. Further, that sample displayed the lowest allelic richness, which compensated for the bias of allelic diversity estimates between samples consisting of different numbers of fish. This observation is consistent with the results of Limborg *et al.* (2009), who inferred a general trend of reduced allelic richness for Baltic Sea samples compared with a sample from the NSEA. Looking to Norwegian fjord samples, all displayed a lower allelic richness than either NSEA or CEL samples (Table 2). It has been concluded before from mtDNA analyses (Debes *et al.*, 2008) and microsatellite analyses (Limborg *et al.*, 2009) that the lower genetic diversity in Baltic sprat compared with NSEA sprat could reflect relatively recent colonization of the region. It is possible too that a similar mechanism is responsible for the lower allelic richness in the Norwegian fjord samples compared with the NSEA samples.

Only small and mostly insignificant genetic differences were observed among the samples of sprat collected in Norwegian fjords, but relatively large and statistically highly significant

genetic differences were observed between all fjord samples and the samples taken in the NSEA and the CEL. Together with the reduced allelic diversity observed in all fjord samples compared with the NSEA and CEL samples, it is our opinion that the data indicate limited connectivity among sea-going sprat (in this context referring to the sprat sampled in the CEL and NSEA) and those found in Norwegian fjords. In turn, this could provide a contributing factor to explaining why there has been a historical decline in Norwegian fjord populations in the same period that there has been a relatively stable catch of sprat in the open seas (ICES, 2011).

A pilot study analysing haemoglobin genetic variation among sprat sampled within Norwegian fjords revealed large genetic differences among some pairs of samples (Nævdal, 1968). In that study, it was concluded that sprat were probably represented by at least two reproductively isolated spawning populations along the Norwegian coastline. However, haemoglobin genotypes have been linked with the growth rate in fish (Imsland *et al.*, 2000), although the relationship between genotype and phenotype is not always clear (Jørstad *et al.*, 2006). Therefore, inferring evolutionary relationships among populations using haemoglobin as the sole genetic marker needs to be done cautiously. Importantly, the haemoglobin analyses reported by Nævdal (1968) did not reveal any geographic pattern among samples of sprat from the Norwegian fjords, and both the samples from the haemoglobin-based and this study were almost exclusively taken during the period when the fishery operated between autumn and early winter as opposed to in the spawning season (spring to early summer). It is therefore possible that any population genetic structure among samples from the Norwegian fjords may have been influenced by migration. This situation may also have been complicated further by temporal variations in the influx of larvae into fjords, resulting from spawning of the sprat populations in the NSEA, the Skagerrak, and the Kattegat. For Atlantic cod off southern Norway, annual fluctuations in the prevailing oceanographic conditions have been demonstrated to cause variations in the local population genetic structure by varying the component of larvae resulting from NSEA spawning and local cod spawning (Knutsen *et al.*, 2004).

Sprat spawning has been documented in Norwegian fjords located on the west coast (Dannevig and Gundersen, 1954; Torstensen, 1984). However, except fish caught in autumn in a fishery targeting primarily age-1 sprat, and the acoustic fjord cruise time series (1968–2008, mainly autumn to early winter), there are limited data on the abundance of sprat within Norwegian fjords over the year. Such demographic information would be invaluable in interpreting the patterns of genetic structure revealed here, specifically because our samples were collected outside the spawning season, when putative populations should be aggregated. Certainly, the fact that only small and mostly statistically insignificant genetic differences were revealed among Norwegian fjord samples, and that all these were highly differentiated from all the other samples analysed here, including the neighbouring NSEA, suggests that there may be considerable gene flow and demographic connection among sprat along the Norwegian coastline.

The only other analysis of sprat made with DNA markers found significant statistical differentiation between sprat sampled in the NSEA and the CEL (Limborg *et al.*, 2009). This contrasts with the results of the present study, in which no genetic differences were observed between sprat sampled from these two areas. It is not possible to elucidate fully the apparent disparity between these two results, but it must be emphasized that the exact location and time of year of the samples taken from these two areas were different between the present study and that of Limborg *et al.* (2009). Mixing different frequencies of partially separated populations on feeding grounds as opposed to distinct populations aggregating during their spawning season may well lead to the disparity between the results of this study and that of Limborg *et al.* (2009).

Marine fish do not display population genetic structure only over extensive distances, but also over short distances (e.g. Knutsen *et al.*, 2003). Nevertheless, one has to urge caution in interpreting small but statistically significant genetic differentiation, because other potential sources of variation, e.g. genotyping errors and non-representative sampling, become relatively more important when the biological signal decreases (Nielsen *et al.*, 2009). Extensive temporal sampling permitted Knutsen *et al.* (2011) to demonstrate that small-scale spatial genetic variation was more important than temporal variation in determining the population genetic structure of cod in the fjords of southern Norway. Hence, to elucidate fully the population genetic structure of sprat among Norwegian fjords and between Norway and the surrounding seas, it is essential that this study be expanded to include extensive temporal sampling. First, within-year temporal sampling, to examine the potential influence of spawning season and the subsequent movements of these fish to and from regions where they are targeted by fisheries in autumn, will be needed. Further, sampling between years, to examine long-term stability and the potential elucidating effects of pulses of larvae from spawning taking place outside the fjords, i.e. drifting from spawning in the NSEA and/or the Skagerrak/Kattegat, will be needed. Finally, expansion of the repertoire of genetic markers for this species may be required to extract the best possible information from the intensive sampling outlined above. Identification of a resource of single-nucleotide polymorphism markers may provide the necessary tools to achieve this. Selection of highly informative markers from larger panels (Glover *et al.*, 2010) may reveal greater genetic differentiation when compared with small suites of microsatellite markers. This may be especially true when identifying markers under natural selection, permitting a greater ability to delineate population structure on an ecological time-scale that is of importance in the active implementation of DNA-based methods in managing fisheries (Waples *et al.*, 2008).

Acknowledgements

We acknowledge the help of Elin E. Danielsen and Åsta B. Stølen in conducting the laboratory work, and crew from the vessels involved in collecting the samples for their support. The work was financed by the Norwegian Ministry of Fisheries and Coastal Affairs.

References

Antao, T., Lopes, A., Lopes, R. J., Beja-Pereira, A., and Luikart, G. 2008. LOSITAN: a workbench to detect molecular adaptation based upon a F_{st}-outlier method. BMC Bioinformatics, 9: 323.

Bekkevold, D., André, C., Dahlgren, T. G., Clausen, L. A. W., Torstensen, E., Mosegaard, H., Carvalho, G. R., *et al.* 2005. Environmental correlates of population differentiation in Atlantic herring. Evolution, 59: 2656–2668.

Dailianis, T., Limborg, M., Hanel, R., Bekkevold, D., Lagnel, J., Magoulas, A., and Tsigenopoulos, C. S. 2008. Characterization of

nine polymorphic microsatellite markers in sprat (*Sprattus sprattus* L.). Molecular Ecology Resources, 8: 861–863.

Dannevig, G., and Gundersen, K. R. 1954. Brislingens gyting. 1. Undersøkelser i Skagerakk og Ryfylke. Av Gunnar Dannevig. 2. Undersøkelser i Hordaland og Sogn Av Kaare R. Gundersen. Fiskeridirektoratets Småskrifter, 3. 19 pp. (in Norwegian).

Debes, P. V., Zachos, F. E., and Hanel, R. 2008. Mitochondrial phylogeography of the European sprat (*Sprattus sprattus* L., Clupeidae) reveals isolated climatically vulnerable populations in the Mediterranean Sea and range expansion in the Northeast Atlantic. Molecular Ecology, 17: 3873–3888.

Dieringer, D., and Schlotterer, C. 2003. Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. Molecular Ecology Notes, 3: 167–169.

Falush, D., Stephens, M., and Pritchard, J. K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics, 164: 1567–1587.

Glover, K. A., Hansen, M. M., Lien, S., Als, T. D., Høyheim, B., and Skaala, Ø. 2010. A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. Genetics, 11: 2. doi: 10.1186/1471-2156-11-2

Goudet, J. 1995. FSTAT (Version 1.2): a computer program to calculate F-statistics. Journal of Heredity, 86: 485–486.

ICES. 2009. Report of the Benchmark Workshop on Short-lived Species (WKSHORT). 31 August–4 September 2009, Bergen, Norway. ICES Document CM 2009/ACOM: 34. 166 pp.

ICES. 2011. Report of the Herring Assessment Working Group for the Area South of 62°N (HAWG), 16–24 March 2011, ICES Headquarters, Copenhagen. ICES Document CM 2011/ACOM: 06.

Imsland, A. K., Foss, A., Stafansson, S. O., and Nævdal, G. 2000. Hemoglobin genotypes of turbot (*Scophthalamus maximus*): consequences for growth and variation in optimal temperature for growth. Fish Physiology and Biochemistry, 23: 75–81.

Jørgensen, H. B. H., Hansen, M. M., Bekkevold, D., Ruzzante, D. E., and Loeschcke, V. 2005. Marine landscapes and population genetic structure of herring (*Clupea harengus* L.) in the Baltic Sea. Molecular Ecology, 14: 3219–3234.

Jørstad, K. E., Karlsen, O., Svåsand, T., and Otterå, H. 2006. Comparison of growth rate among different protein genotypes in Atlantic cod, *Gadus morhua*, under farming conditions. ICES Journal of Marine Science, 63: 235–245.

Jørstad, K. E., and Nævdal, G. 1981. Enzyme polymorphism of sprat from Norwegian waters – preliminary results. ICES Document CM 1981/H: 65. 9 pp.

Knutsen, H., André, C., Jorde, P. E., Skogen, M. D., Thuróczy, E., and Stenseth, N. Ch. 2004. Transport of North Sea cod larvae into the Skagerrak coastal populations. Proceedings of the Royal Society of London, Series B, 271: 1337–1344.

Knutsen, H., Jorde, P. E., André, C., and Stenseth, N. Ch. 2003. Fine-scaled geographical population structuring in a highly mobile marine species: the Atlantic cod. Molecular Ecology, 12: 385–394.

Knutsen, H., Olsen, E. M., Jorde, P. E., Espeland, S. H., André, C., and Stenseth, N. Ch. 2011. Are low but statistically significant levels of genetic differentiation in marine fishes "biologically meaningful"? A case study of coastal Atlantic cod. Molecular Ecology, 20: 768–783.

Limborg, M. T., Pedersen, J. S., Hemmer-Hansen, J., Tomkiewicz, J., and Bekkevold, D. 2009. Genetic population structure of European sprat *Sprattus sprattus*: differentiation across a steep environmental gradient in a small pelagic fish. Marine Ecology Progress Series, 379: 213–224.

Nævdal, G. 1968. Studies on haemoglobins and serum proteins in sprat from Norwegian waters. Fiskeridirektoratets Skrifter, Serie Havundersøkelser, 14: 160–182.

Neilson, E. E., Hansen, M. M., and Meldrup, D. 2006. Evidence of microsatellite hitch-hiking selection in Atlantic cod (*Gadus morhua* L.): implications for inferring population structure in non-model organisms. Molecular Ecology, 15: 3219–3229.

Nielsen, E. E., Wright, P. J., Hemmer-Hansen, J., Poulsen, N. A., Gibb, I. M., and Meldrup, D. 2009. Microgeographical population structure of cod *Gadus morhua* in the North Sea and west of Scotland: the role of sampling loci and individuals. Marine Ecology Progress Series, 376: 213–225.

Pampoulie, C., Ruzzante, D. E., Chosson, V., Jörunsdóttir, D. D., Taylor, L., Torsteinsson, V., Daníelsdóttir, A. K., *et al.* 2006. The genetic structure of Atlantic cod (*Gadus morhua*) around Iceland: the pan I locus and tagging experiments. Canadian Journal of Fisheries and Aquatic Sciences, 63: 2660–2674.

Pritchard, J. K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics, 155: 945–959.

Rice, W. R. 1989. Analyzing tables of statistical tests. Evolution, 43: 223–225.

Raymond, M., and Rousset, F. 1995. GENEPOP (Version 1.2) – Population-genetics software for exact tests and ecumenicism. Journal of Heredity, 86: 248–249.

Takezaki, N., Rzhetsky, A., and Nei, M. 1995. Phylogenetic test of the molecular clock and linearized trees. Molecular Biology and Evolution, 12: 823–833.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology and Evolution, 24: 1596–1599.

Torstensen, E. 1984. Sprat spawning in two fjord areas of western Norway in 1982 and 1983. ICES Document CM 1984/H: 41. 16 pp.

Waples, R., Punt, A. E., and Cope, J. M. 2008. Integrating genetic data into management of marine resources: how can we do it better? Fish and Fisheries, 9: 423–449.

Zardoya, R., Castilho, R., Grande, C., Favre-Krey, L., Caetano, S., Marcato, S., Krey, G., *et al.* 2004. Differential population structuring of two closely related fish species, the mackerel (*Scomber scombrus*) and the chub mackerel (*Scomber japonicus*), in the Mediterranean Sea. Molecular Ecology, 13: 1785–1798.

Chapter 5

Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges

Published in *Molecular Ecology Resources*

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11 (Suppl. 1),** 123–136 doi: 10.1111/j.1755-0998.2010.02943.x

ANALYTICAL APPROACHES

Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges

S. J. HELYAR,* J. HEMMER-HANSEN,† D. BEKKEVOLD,† M. I. TAYLOR,* R. OGDEN,‡ M. T. LIMBORG,† A. CARIANI,§ G. E. MAES,¶ E. DIOPERE,¶ G. R. CARVALHO* and E. E. NIELSEN†
Molecular Ecology and Fisheries Genetics Laboratory (MEFGL), School of Biological Sciences, University of Bangor, Environment Centre Wales, Bangor, Gwynedd LL57 2UW, UK, †National Institute of Aquatic Resources, Technical University of Denmark, Vejlsøvej 39, DK-8600 Silkeborg, Denmark, ‡TRACE Wildlife Forensics Network, Royal Zoological Society of Scotland, Edinburgh EH12 6TS, UK, §Molecular Genetics for Environmental & Fishery Resources Laboratory – GenMAP, Interdepartmental Centre for Research in Environmental Sciences, University of Bologna, Ravenna 163-48100, Italy, ¶Laboratory of Animal Diversity and Systematics, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium

Abstract

Recent improvements in the speed, cost and accuracy of next generation sequencing are revolutionizing the discovery of single nucleotide polymorphisms (SNPs). SNPs are increasingly being used as an addition to the molecular ecology toolkit in nonmodel organisms, but their efficient use remains challenging. Here, we discuss common issues when employing SNP markers, including the high numbers of markers typically employed, the effects of ascertainment bias and the inclusion of nonneutral loci in a marker panel. We provide a critique of considerations specifically associated with the application and population genetic analysis of SNPs in nonmodel taxa, focusing specifically on some of the most commonly applied methods.

Keywords: ascertainment bias, nonneutral loci, outlier detection, population genomics, population structure, software

Received 14 July 2010; revision received 23 September 2010; accepted 27 September 2010

Introduction

Recent improvements in the speed, cost and accuracy of next generation sequencing (NGS) and advances in the accompanying bioinformatic tools are revolutionizing the opportunities for generating genetic resources in nonmodel organisms. In turn, this is driving a shift from anonymous markers such as microsatellites to direct analyses of sequence variation including single nucleotide polymorphisms (SNPs). This shift has evolved from the initial uptake of such markers in humans and other commercially important species, to their application in a wide range of nonmodel species.

SNPs are attractive markers for many reasons (for reviews see Brumfield *et al.* 2003; Morin *et al.* 2004), including the availability of high numbers of annotated markers, low-scoring error rates, relative ease of calibration among laboratories compared to length-based markers and the associated ability to assemble combined temporal and spatial data sets from multiple laboratories. Additionally, the potential for high-throughput genotyping improved genotyping results for poor quality samples [such as historical, noninvasive or degraded samples (Morin & McCarthy 2007; Smith *et al.* 2011)], a simple mutation model, and the ability to examine both neutral variation and regions under selection offers unparalleled scope for expansive screening of genomes and large sample sizes from natural populations. Although several early studies questioned the advantage of SNPs over neutral markers such as microsatellites (e.g. Rosenberg *et al.* 2003), more recent studies have shown that SNPs are also showing promise as highly informative markers, as many studies with access to very large numbers of SNPs (mainly human) have shown that a small fraction of the SNPs have a very high information content for population structure analysis (e.g. Lao *et al.* 2006; Paschou *et al.* 2007), outperforming microsatellites (Liu *et al.* 2005). Despite microsatellites typically displaying far greater allelic diversity per locus, individual SNPs can segregate strongly among populations (Freamo *et al.* 2011; Karlsson *et al.* 2011).

Correspondence: S.J. Helyar, Fax: 01248 370731;
E-mail: s.helyar@bangor.ac.uk

Although SNPs are increasingly being used as an addition to the molecular ecology toolkit, their use as a standard tool in nonmodel organisms remains challenging, with debate over how to utilize them most efficiently. A recent study by Garvin *et al.* (2010) reviewed the technical aspects of SNP discovery and genotyping, but there are also challenges associated with the analysis of SNP data. These concerns vary depending on the questions being addressed: some specific issues have been covered in other papers (e.g. parentage assignment, Anderson & Garza 2006; Hauser *et al.* 2011; power assessment, Morin *et al.* 2009; development of linkage maps, Ball *et al.* 2010 and relatedness, Krawczak 1999). However, an overview of the considerations specifically associated with the application of SNPs and their appropriate analysis in population genetic studies of nonmodel organisms appears timely. We focus specifically on some of the most commonly applied methods and first discuss the challenges common to all analyses; problems arising from the dramatic increase in the number of markers that are available, the effects of ascertainment bias and the inclusion of nonneutral loci in a marker panel.

Number of loci

Using SNP data to analyse population structure is theoretically straightforward, but until recently a major obstacle was the identification of software that could handle large data sets. However, for many of the standard analyses, such as basic descriptive statistics, authors have modified their software to accept several thousand loci (see Table 1). Nevertheless, many packages are still limited by either the number of loci or the sum of individuals × loci that can be analysed. Additional problems may also arise when using some analytical methods that are computationally intensive, such as Bayesian MCMC methods. While such software may accept very large data sets, the time taken for a standard desktop computer to conduct the analysis may be prohibitive.

Ascertainment bias

Ascertainment bias is the systematic deviation from the expected allele frequency distribution that occurs because of the sampling processes used to find (ascertain) marker loci. In SNPs, this may occur as the markers are generally identified in a small panel of individuals from part of the species' range (ascertainment width). Likewise, only SNPs occurring more than a predefined number (k) of times in the ascertainment sample are included (ascertainment depth). When these SNPs are then genotyped on a larger sample of individuals, an 'ascertainment bias' is introduced (Nielsen 2000; Albrechtsen *et al.* 2010). Because of the small size of the ascertainment panel (compared to the population), the probability that a SNP is identified in this panel is a function of its minor allele frequency (MAF), i.e. SNPs with a very low MAF are less likely to be discovered than those with a higher MAF.

Ascertainment bias may compromise analyses based on diversity measures, for example, any statistical measure that relies on allele frequency may be affected. Because there is a bias towards not sampling rare SNPs, the average diversity of polymorphic sites is overestimated, while the average diversity across all sites is underestimated. This may lead to a bias in estimates of nucleotide diversity, population size, demographic changes, linkage disequilibrium, selective sweeps and inferences of population structure (Nielsen 2000; Schlötterer & Harr 2002; Akey *et al.* 2003; Nielsen & Signorovitch 2003; Marth *et al.* 2004; Rosenblum & Novembre 2007; Storz & Kelly 2008; Guillot & Foll 2009; Chen *et al.* 2010; Moragues *et al.* 2010). The size and direction of the bias depend on the sampling strategy used for the ascertainment panel; for example, studies on both humans and *Drosophila* suggest that genetic diversity will be underestimated if individuals from the ancestral population range are not included in the ascertainment panel (Schlötterer & Harr 2002; Romero *et al.* 2009). However, a panel based on purely ancestral (African) *Drosophila* did not underestimate the diversity in the European populations. Moreover, a study by Rosenblum & Novembre (2007) that examined a spatially structured population of lizards found that choosing individuals at random from across the geographical range minimized the resulting bias. However, some studies with small ascertainment panels are not addressing these issues (e.g. Kerstens *et al.* 2009; Li *et al.* 2010).

Three main approaches have been used to address and correct for ascertainment bias in studies of natural populations: (i) the application of more robust methods, such as those based on haplotype structure (e.g. Sabeti *et al.* 2007, however, this requires a full genome as reference), (ii) the simulation of data based on the ascertainment process to derive appropriate critical values and confidence intervals taking the ascertainment into account (e.g. Carlson *et al.* 2004; Voight *et al.* 2006) and (iii) the direct correction of the statistical estimators and statistics using specific models (e.g. Nielsen 2000; Wakeley *et al.* 2001; Nielsen & Signorovitch 2003; Polanski & Kimmel 2003; Marth *et al.* 2004; Nielsen *et al.* 2004;. Also see Table 1). However, a major restriction is that the correction of the allele frequency spectrum restricts downstream analyses to corrected summary statistic data (allele frequencies) with the loss of the observed individual genotypes that are needed for many applications (e.g. determining population structure, individual

Table 1 Computer software used for the most common aspects of population genetics

Programme	Functions	Maximum number of loci	Maximum number of individuals	Reference and web address
PEAS v1	Multiple data manipulation and summary statistics	None	None	Xu *et al*. (2010). http://www.picb.ac.cn/~xushua/index.files/Download_PEAS.htm Data manipulation includes file conversion for other population genetics programmes
SNPator	Multiple data manipulation and summary statistics	None	None	Morcillo-Suarez *et al*. (2008). http://www.snpator.org/public/downloads/aRamirez/tajimasDCorrector/
POPGENE	Multiple summary statistics	1000	1400 pops/150 groups	http://www.ualberta.ca/~fyeh/
Arlequin 3.5*	Multiple summary statistics	None	None	Excoffier & Lischer (2010). http://cmpg.unibe.ch/software/arlequin35/
Genepop v4	Multiple summary statistics	None	None	Rousset (2008). http://kimura.univ-montp2.fr/~rousset/Genepop.htm
popgen†	Multiple	None	None	http://mathgen.stats.ox.ac.uk/software.html
FSTAT2.9.4	Multiple summary statistics	10 000	200	Goudet (1995). http://www2.unil.ch/popgen/softwares/fstat2.9.4_10kloc_9all_200pops.zip
HIERFSTAT†	F-statistics	None	None	Goudet (2005). http://www.unil.ch/popgen/softwares/hierfstat.htm.
GenAlEx6.4	Multiple summary statistics	127 or 8192‡	65 500	Peakall and Smouse (2006). http://www.anu.edu.au/BoZo/GenAlEx/index.php
Genetix4.05	Multiple	None	None	http://www.genetix.univ-montp2.fr/genetix/genetix.htm
AscB†	Correction for Ascertainment Bias	None	None	Guillot & Foll (2009) http://www2.imm.dtu.dk/~gigu/AscB/
trueFS	Correction for Ascertainment Bias	None	None	Nielsen *et al*. (2004). http://people.binf.ku.dk/rasmus/webpage/truefs.html
Plink1.07§	Multiple	None	None	Purcell *et al*. (2007). http://pngu.mgh.harvard.edu/purcell/plink/
DetSel	Outlier locus detection	None	None	Vitalis *et al*. (2003). http://www.genetix.univ-montp2.fr/detsel.html
FDIST2	Outlier locus detection	None	None	Detection of loci under selection from hierarchical F-statistics, implemented in Arlequin (see above)
BAYESFST¶	Outlier locus detection	None	None	Beaumont & Balding (2004). http://www.reading.ac.uk/Statistics/genetics/software.html
LOSITAN	Outlier locus detection	None	None	Antao *et al*. (2008) http://popgen.eu/soft/lositan/
BayeScan	Outlier locus detection	None	None	Foll & Gaggiotti (2008). http://www-leca.ujf-grenoble.fr/logiciels.htm
matSAM v2	Outlier locus detection	None	None	Joost *et al*. (2008). http://www.econogene.eu/software/sam/
Structure 2.3.3	(Spatial) Genetic Structure	The maximum data set size around 100 million genotypes (loci × ind.)**††		Pritchard *et al*. (2000). http://pritch.bsd.uchicago.edu/structure.html
PCAGEN	Genetic Structure	50	5000 ind. 500 pops	http://www2.unil.ch/popgen/softwares/pcagen.htm

Table 1 Continued

Programme	Functions	Maximum number of loci	Maximum number of individuals	Reference and web address
adegenet†	Genetic Structure	None	None	Jombart (2008). http://adegenet.r-forge.r-project.org/
Geneland†	Spatial Genetic Structure	None**	None**	Guillot & Santos (2009). http://www2.imm.dtu.dk/~gigu/Geneland/
TESS 2.3	Spatial Genetic Structure	None**	None**	Chen *et al.* (2007). http://membres-timc.imag.fr/Olivier.Francois/tess.html
BAPS	Genetic Structure	None**	None**	Corander *et al.* (2008). http://web.abo.fi/fak/mnf/mate/jc/smack_software_eng.html
GESTE	Genetic Structure	None**	None**	Foll & Gaggiotti (2006). http://www-leca.ujf-grenoble.fr/logiciels.htm
GeneClass2	Assignment	None**	None**	Piry *et al.* (2004). http://www.ensam.inra.fr/URLB/GeneClass2/Setup.htm
WHICHLOCI	Locus selection	None**	None	Banks *et al.* (2003). http://www.bml.ucdavis.edu/whichloci.htm
GAFS 1.1	Locus selection	None**	None	Topchy *et al.* (2004). http://www.fw.msu.edu/~scribne3/molecularecology/programs.htm
BELS	Locus selection	None**	None	Bromaghin (2008). http://alaska.fws.gov/fisheries/biometrics/programs.htm

*Although Arlequin is not an R package, the latest version interfaces with R to produce the graphs.
†An R package. Additional packages may be found at http://cran.r-project.org/web/views/Genetics.html
‡The number of loci is dependant of the version of excel that you use, for pre-2007 as the number of columns in Excel was 256, but this has increased in Excel 2007 to 16 384 columns. For versions of GenAlEx 6.3 onwards, users are given the choice of installing either GenAlEx6.3.xla or GenAlEx 6.3 for 2007.xla. Both versions will run in Excel 2007, but to take advantage of full compatibility with Excel 2007 you should instal the Excel 2007 specific option.
§Extensible with via R function plug-ins.
¶R scripts also available on the website.
**While the are no physical constraints on the numbers of loci or individuals that can be submitted to this programme, the number of permutations that may needed for the computation of some options may make the calculation prohibitive on a standard desktop computer.
††The authors suggest reducing the data set for the exploratory analysis. Additionally, for large data sets, the default settings for BURNIN and NUMREPS can be reduced, without affecting the accuracy.

assignment, multilocus heterozygosity estimates, mixed stock analysis).

The generation of more and longer reads will eventually lead to the next step in SNP genotyping, where individuals are directly (single track) sequenced, followed by a high-confidence assembly and phasing of sequence reads [using for example; Phase (Stephens *et al.* 2001), FastPhase (Scheet & Stephens 2006), Shape-IT (Delaneau *et al.* 2008)]. Alternatively, the genotyping-by-sequencing approach, used for instance in RAD sequencing, combines the power of high throughput sequencing and large-scale polymorphism genotyping in one step (for a limited number of individuals) significantly reducing the problem of ascertainment bias (Baird *et al.* 2008; Hohenlohe *et al.* 2010).

However, as NGS data is likely to remain the basis of SNP development for the foreseeable future (see conclusions), and considering the inherent properties of newly developed SNPs, ascertainment bias is likely to remain a problem in the near future and may lead to incorrect population genetic inferences. Consequently, attempts must be made both to minimize the effects by careful design of the ascertainment panel. This can be achieved by the geographical sampling of multiple individuals, the tagging of individuals used in the sequencing for later genotype/haplotype reconstructions and a sufficient sequencing depth for *in silico* frequency spectra to be assessed before final SNP genotyping (for instance, by combining long (454 Roche) and short (ABI, Illumina) read sequencing runs for reference assembly and SNP discovery,

respectively). Accounting for the bias in the resulting data with the use of up to date statistical and simulation/modelling tools will allow the robustness of results to be assessed, despite assumption violations (Balzer *et al.* 2010). However, as explored in more detail in the following sections, ascertainment bias need not always pose a problem.

Nonneutral loci

The availability of thousands of genetic markers reinforces the need for careful evaluation of the markers used for a specific population genetic study, as markers in genic and nongenic regions may generally differ with respect to basic properties such as levels of variation and population differentiation, which will affect the outcome of downstream analyses. In genome-wide association (GWA) studies in humans, it has been found that SNPs in genic regions are more likely to display signatures of both positive and negative selection than those in nongenic regions (Barreiro *et al.* 2008; Coop *et al.* 2009) and that genetic variation is generally lower in gene-rich regions (Cai *et al.* 2009). While the degree that these findings apply to nonmodel organisms remains unknown, they do indicate that markers situated in or close to genes may not provide a representative picture of genome-wide effects of neutral evolutionary forces. Additionally, genomes contain gene regulatory networks (GRNs) that are highly conserved regions within the noncoding DNA (Davidson *et al.* 2002; Woolfe *et al.* 2005); this implies both that these regions will not be identified by transcriptome sequencing and also that there are sections of noncoding DNA that are under selection.

SNPs represent the most widespread type of sequence variation in genomes, and the combination of the continuing decrease in costs for NGS and new efficient methodologies, such as RAD-tag sequencing (e.g. Miller *et al.* 2007; Baird *et al.* 2008) and RRS (reduced representation sequencing—e.g. Castano-Sanchez *et al.* 2009), is showing great promise for fast, efficient SNP detection in nonmodel species. While there are methods that can preferentially target noncoding regions (e.g. EPIC markers, Palumbi & Baker 1994), there are also increasing expressed sequence tag (EST) resources available for many taxa, increasing the likelihood that many SNP loci that are being developed will be located either within or very close to coding regions. However, it is now thought that animal genomes are pervasively transcribed (Ponting *et al.* 2009) with a large number of noncoding transcripts being polyadenylated, which will therefore be included in EST collections. Consequently, the representation of the genome might be larger and have fewer constraints on sequence variability than previously thought.

For some applications, this potential bias in genome coverage has been highlighted as an advantage, if for example, the aim is to identify candidate genes under selection (Bonin 2008; Brieuc & Naish 2011; Hemmer-Hansen *et al.* 2011 and also see the discussion in the section 'Detection of Outliers' below). However, issues could arise if the purpose of a study is to make general inferences about neutral evolutionary processes, such as genetic drift and gene flow. In such cases, markers under selection should be removed prior to analyses (Beaumont & Nichols 1996), as they may bias results significantly (see also discussion in Laval *et al.* 2010 and below). On the other hand, markers under selection could be exploited for specific purposes, such as investigating population structure on ecological rather than evolutionary timescales (Waples & Gaggiotti 2006), and for increasing the power for assigning individuals to populations of origin (Nielsen *et al.* 2009b).

With these caveats in mind, we now review the application of the most common analytical methods in population genetics to SNP data, paying special attention to the significant issues described, particularly how ascertainment bias and nonneutral loci affect analyses and how such effects can be addressed. Finally, we highlight salient priorities for further research in the integration of SNPs into molecular ecology.

Population genetic data analyses

Measures of genetic differentiation and population structure

With the ever increasing opportunities for SNP mining in nonmodel species, it is becoming increasingly evident that the apparent shortcomings of individual SNPs to detect population structure compared to microsatellites (Rosenberg *et al.* 2003) can be overcome by the relative ease with which large numbers of SNP markers can be developed and screened. The statistical power to detect population structure is related to the total number of alleles examined, and the discriminatory power of ~100 (neutral) SNPs is very roughly equivalent to 10–20 microsatellites (Kalinowski 2002). Moreover, the most informative SNP markers (i.e. those that show the greatest allele frequency variation among populations) in a panel may rival (or even exceed) the average information content of microsatellite markers (e.g. Liu *et al.* 2005; Smith *et al.* 2007). Using SNP markers to investigate population structure is theoretically straightforward, and most standard population genetic software packages allow for inclusion of large numbers of loci. However, there are also practical considerations, as some (especially Bayesian) methods are computationally intensive and may have problems handling very large data sets.

Wright's *F*-statistics are arguably the most commonly used descriptive statistics in population and evolutionary genetics (Wright 1931). As their original development, many related statistics have been described either as improvements or for specific applications, for example, for microsatellite data (G_{ST}, θ, and R_{ST}), sequence data (Φ_{ST}) and for quantitative traits (Q_{ST}) (see Holsinger & Weir 2009 for a review). One issue that has caused much debate is how to compare diversity estimates among markers, with much focus on the effect of differing mutation rates and levels of heterozygosity between highly polymorphic markers, such as microsatellites, and less variable markers, such as allozymes and SNPs (Waples & Gaggiotti 2006; Allendorf & Luikart 2007). In 2005, Hedrick proposed the new statistic G'_{ST} to provide a measure of differentiation that allows comparison among loci with different levels of genetic variation, such as among microsatellites, or between different marker types, such as allozymes/SNPs and microsatellites; measures such as this and the more recent D_{EST} (Jost 2008) are increasingly being used (e.g. De Carvalho *et al.* 2010; White *et al.* 2010). However, G'_{ST} has also been criticized as uninformative when migration is not expected to be negligible (Ryman & Leimar 2008). Mutation rates are in general considerably lower for SNPs than for microsatellites (Foll & Gaggiotti 2008; Excoffier *et al.* 2009), and more importantly while the expected locus-specific heterozygosity may reach more than 0.95 for a microsatellite marker, the maximum expected heterozygosity that can be reached by a bi-allelic SNP is 0.5. Such constraints mean that single locus F_{ST} estimates derived from SNP markers are likely to be more comparable than those derived from microsatellite loci. Many of the most frequently used programmes for calculating F_{ST} and related statistics have recently extended their capacity for numbers of loci and samples (details shown in Table 1).

Within human genetics, large-scale GWA studies are increasingly focusing on the population genetics of the samples, as unidentified structure may lead to spurious associations between traits and markers/genes. While such factors have enhanced the development of some SNP-specific software, such as Plink (Purcell *et al.* 2007), it has yet to be seen how applicable these are to more traditional population genetic approaches in nonmodel organisms.

In nonmodel species, global and pairwise F_{ST} values are typically estimated over all loci; as all markers are assumed to be effectively neutral, there should not be any major inconsistencies between loci. However, when loci are potentially under different selective pressures the estimates may be different for each locus, requiring per locus estimates. Xu *et al.* (2009) proposed a new measure of population structure specifically for SNPs. It is based on the c parameter (Nicholson *et al.* 2002), which is population-specific and measures the differentiation of the population from the common ancestral population. In contrast, the new measure C is an index of the overall levels of population structure across populations. Extensive simulations in Xu *et al.* (2009) show that C takes into account ascertainment bias and correlates well with Wright's F_{ST}. The correlation increases with increasing information (more SNPs and/or more subpopulations in the samples).

Clustering algorithms such as Bayesian MCMC clustering approaches are frequently utilized in genetic analyses. These methods define populations by minimizing departures from Hardy–Weinberg and maximizing linkage equilibrium. Clustering analyses can be performed independently of spatial information or be linked analytically to spatial and/or environmental parameters; the latter commonly termed 'landscape genetics' (Guillot *et al.* 2009). Genetic clustering and analyses of spatial structure can be based on neutral marker variation, on markers under selection or on a combination, with the last of these commonly being of particular interest in many EST-derived SNP approaches. User-friendly software for conducting such analyses includes Structure (Pritchard *et al.* 2000), BAPS (Corander *et al.* 2008), GESTE (Foll & Gaggiotti 2008) and Geneland (Guillot *et al.* 2005), the current versions of which all allow for the inclusion of larger numbers of loci (see Table 1). However, the assumptions of no linkage disequilibrium between markers common to many of these applications are likely to be violated with denser SNP coverage/representation across chromosomal regions, although some applications do allow the inclusion of linkage information (Falush *et al.* 2003). Including a relatively low number of markers in linkage disequilibrium is not likely to bias estimates of population differentiation, but may lead to overestimates of clusters (Kaeuffer *et al.* 2007). However, the effects of including markers with different levels of linkage disequilibrium on estimates of cluster numbers and divergence are not well described. As an alternative to Bayesian clustering, principal component analysis (PCA) and related approaches have been applied in several SNP studies of human population structure (Patterson *et al.* 2006). An advantage of PCA-based approaches, compared to Bayesian methods, is that PCA can be performed quickly on desktop computers. PCA approaches also facilitate the identification of subsets of markers that effectively describe differences among populations (Paschou *et al.* 2007), and it has even been argued that PCA outperforms Bayesian methods for inferring population structure when many loci are available and the structure is subtle (Reeves &

Richards 2009). However, PCA methods are sensitive to missing data and sampling effects, especially for species and populations with continuous distributions (Novembre & Stephens 2008), which can limit inference about underlying historical and demographic processes [although ways of circumventing these problems have been proposed for SNP data (Paschou *et al.* 2007)].

SNP-based estimates of population structure are potentially affected by ascertainment bias if the SNP panel used was developed for populations (or species) other than those analysed (Nielsen 2000). Nonetheless, few statistical assessments of the effect of ascertainment bias on fundamental measures such as F_{ST} estimates have been reported (although see Schlötterer & Harr 2002; Albrechtsen *et al.* 2010; and Moragues *et al.* 2010). Including information for loci either under directional or balancing selection themselves or loci tightly linked to regions under selection leads to violation of assumptions for most neutral population genetic models and may cause erroneous inference about population demographic parameters, such as rates of genetic drift and migration between individual demes. Several reports of population structure based on presumably neutral marker information are likely to (unknowingly) have incorporated nonneutral markers (Nielsen *et al.* 2006). In weakly structured species, the effect of just a few loci on overall patterns could be significant, but provided selected loci make up only a small proportion of the total marker number, biological inference is not generally expected to be severely biased (Luikart *et al.* 2003). Nonetheless, with SNP markers often developed from transcriptomic sequencing, the dramatic increase in genome coverage implies that some proportion of the markers are likely to be linked to genes/regions under selection, making it of paramount importance to test for marker 'neutrality' prior to exploring population structure (for example, by using outlier tests as outlined in the section below). Studies that combine information from neutral and nonneutral markers in analyses of population structure and estimation of demographic parameters are still scarce for nonmodel organisms (for examples see Gaggiotti *et al.* 2009; Nielsen *et al.* 2009a), and there is a need for development of analytical tools that allow integration across marker classes (Guillot *et al.* 2009).

Detection of outliers

The search for signatures of selection in molecular data has a long tradition in evolutionary biology. Most methods rely on the concept of genetic hitch-hiking (Maynard Smith & Haigh 1974), where a marker is linked to a site under selection, and although not the target of selection, the 'hitch-hiking' marker fails to display patterns of neutrality. For molecular markers, the methods to detect outlier loci can be divided into two broad categories, the first based on linkage disequilibrium between markers, and the second based on differences in levels of genetic variation and levels of genetic divergence between samples (see also Vasemägi & Primmer 2005).

Genome scan approaches (see Luikart *et al.* 2003 and Storz 2005 for reviews) have now been applied to an increasing number of nonmodel organisms (e.g. Anderson *et al.* 2005; Bonin *et al.* 2006; Hayes *et al.* 2007; Eveno *et al.* 2008; Moen *et al.* 2008; Namroud *et al.* 2008; Nielsen *et al.* 2009a), and this has generated insight into the pros and cons to the various approaches for detecting markers under selection in the wild.

Many nonmodel species still have little or no genomic resources, and the location of SNPs within the genome is therefore often unknown, rendering methods relying on detailed analyses of linkage disequilibrium unfeasible. Methods based on comparisons of genetic variation in random sets of markers have been developed both for microsatellites (Schlötterer 2002; Kauer *et al.* 2003; Marshall & Weiss 2006) and SNP-based haplotypes (Voight *et al.* 2006; Sabeti *et al.* 2007); however, these do not seem to be relevant for a relatively limited number of SNPs without genomic information. In contrast, many methods based on comparisons of levels of genetic divergence between samples can be applied to markers where information about genomic location is missing. Hence, these methods appear better suited for studies in nonmodel species.

Most methods based on comparisons of divergence among samples are based on the original Lewontin–Krakauer test, which compares single locus estimates of F_{ST} to an expected neutral distribution of F_{ST} (Lewontin & Krakauer 1973). The original Lewontin–Krakauer test is now rarely used, mainly because of concerns over its performance when allele frequencies are correlated between samples leading to an increased number of false positives (Robertson 1975; Beaumont 2005). However, several closely related methods have been proposed to overcome the shortcomings of the original approach. With a very large number of markers, it may be possible simply to estimate the expected distribution of F_{ST} from the markers themselves (e.g. Akey *et al.* 2002), but for most nonmodel organisms, the available number of markers is too limited and simulations must be used to generate the neutral distribution. In these cases, the model used for the simulations is crucially important, as it will effect the identification of outlier loci. For instance, Vitalis *et al.* (2001, 2003) developed a method (implemented in DetSel) based on pairwise population comparisons of individual locus F_{ST} to a simulated distribution of F_{ST} generated under a model of two fully isolated populations descended from a common ancestral population. Beaumont & Nichols (1996) developed FDIST2, which

uses a classical island model to generate the expected neutral distribution of F_{ST} estimates. While these methods remove the need to directly use genotyped markers as the baseline, they do so indirectly by using the estimated overall F_{ST} as a starting point for simulations. Thus, including loci under selection in the initial F_{ST} estimate may generate a bias in the simulated distribution. Additionally, the models used for the simulation of data in the two methods are unlikely to match most natural situations, because many populations are significantly connected through asymmetrical patterns of gene flow. The two limitations above have been addressed in later Bayesian methods based on logistic regression models of locus and population effects on F_{ST}. Both BAYESFST (Beaumont & Balding 2004) and BayeScan (Foll & Gaggiotti 2008) allow F_{ST} to vary between populations and identify loci potentially under selection through estimates of locus effects on F_{ST}. The two methods are based on the same basic regression model, but differ in the way that the effect of selection is inferred. While BAYESFST does not conduct a formal statistical test, BayeScan uses a likelihood ratio test to assess the most likely of the two alternative models (no effect of selection vs. effect of selection). Both programmes have been widely applied, but they have also recently been found to be vulnerable to complex population structure scenarios, such as when populations are hierarchically structured, leading to correlated allele frequencies among samples (Excoffier *et al.* 2009). A modified, hierarchical, version of FDIST2 implemented in the Arlequin 3.5 software may be more appropriate for such situations (Excoffier *et al.* 2009). The implementation of a hierarchical island model results in higher variance between simulated neutral loci and thus leads to a more conservative estimate of the number of outlier loci (Excoffier *et al.* 2009). It seems inevitable that the lower false-positive rate comes at the expense of a higher false-negative rate; however, the method has so far only been evaluated with simulated neutral loci, focusing on the discovery of false positives, rather than the power for discovering true positives.

While the Bayesian methods may be relatively powerful for detecting directional selection, they have low power for detecting loci under balancing selection, particularly for SNP applications (Beaumont & Balding 2004; Foll & Gaggiotti 2008). This may be problematic in situations with low levels of population structure, when the power for detecting directional selection could be substantially higher than the power for discriminating between loci under balancing selection and loci evolving under neutrality.

In general, a low number of samples also substantially reduces the statistical power of these methods (Foll & Gaggiotti 2008), meaning that pairwise comparisons (e.g. between populations under different environmental forcing) will detect only extreme outlier loci, and many potential candidate loci may be missed. In contrast, too many samples could also bias results, particularly if allele frequencies are correlated among samples, resulting in increased false-positive rates (Excoffier *et al.* 2009). This bias could be reduced through analysing balanced subsets of samples, i.e. using a similar number of samples from each of a number of populations or groups of populations identified through other approaches, such as clustering methods. Thus, a balanced design could minimize effects from complex population structure not easily handled by many current methods. Furthermore, it is possible to evaluate the effect of study design by running several tests on different subsets of samples.

The genetic resource originally used for developing the genetic markers can impact results of outlier detection approaches in several ways. For instance, it must be remembered that in current studies of nonmodel organisms, markers will often mainly be linked to the variation in coding (and expressed) parts of the genome (see section on nonneutral loci). Although the effects of such an ascertainment strategy on genome scans have yet to be assessed, in some approaches, these markers will be used to generate the expected 'neutral' distribution of F_{ST} values. However, if this baseline is biased, then results may not truly reflect the proportion of loci under selection. Further biases may be introduced through ascertainment bias (see introduction and discussion in Nielsen *et al.* 2009b). In addition, loci in linkage disequilibrium could bias results by introducing biased genome coverage among the markers, for instance biassing F_{ST} through physical linkage of loci displaying elevated or lowered levels of structuring.

Although the aforementioned methods have their limitations, they have all been developed to handle relatively large data sets and they are very useful for providing a general overview of the data at hand. Again, the important thing is to have clarity in the question that is being addressed. If the goal is to identify sets of markers with high discriminatory power between different populations/groups of populations, then in principle it does not matter if a detected outlier is truly subject to selection, or if it is a false positive, provided that the signal is temporally stable. In this case, the outlier detection can be viewed as an explorative and preliminary exercise supporting downstream analyses. However, if evolutionary or demographic processes are being investigated, the inclusion of loci under selection may influence results significantly and careful attention should be paid to the design of the scan for outlier loci.

Power analysis

Several population genetic applications, such as conservation management, product traceability and forensic

genetic analysis, involve the assignment of individuals, or collections of individuals, to population of origin based on their (multilocus) genotypes (Manel *et al.* 2005). Here, the inclusion of markers exhibiting evidence for diversifying selection need not violate assumptions and can dramatically increase assignment success, at least if all (or most) reference populations are represented in the baselines against which samples are compared. Analyses combining marker types should, however, be accompanied by simulations of how potential sampling effects could influence assignment (see Anderson 2010). Likewise, inclusion of nonneutral markers may be advantageous when attempting to estimate genetic admixture of individuals or populations.

For applications such as individual assignment (IA), there are many advantages (for example, the reduction in costs, time and computational demands) in using a reduced panel of markers that have been identified as maximizing the power available. For example, selection of breed-informative SNP markers for IA in cattle enabled a reduction in panel size from 54 000 to 200 SNPs with negligible loss of assignment power in twelve European cattle breeds (Wilkinson *et al.*, pers.com). However, a marker panel that has been reduced for this purpose is not suitable for many standard population genetic analyses because of the bias introduced through the high grading of markers that segregate among target populations (Waples 2010).

Identifying loci with maximum power. Not all genotyped loci are necessary for increasing assignment power. Loci may have high-genotyping error rates, be noninformative with little discriminatory power or be strongly correlated (linked) with other markers, thereby yielding redundant information. For some purposes, it may be desirable to create 'minimal panels with maximum power', for example; panels for assigning individuals to major groups, or very specific panels for discriminating between two alternative hypotheses in relation to individual assignment. The selection of loci to form SNP panels for assignment will be driven by the complexity of the assignment question involved. A biallelic marker will only ever be able to segregate two populations; therefore, multiple SNPs will be needed for IA when there are multiple candidate source populations. By assessing assignment power at the level of the individual SNP, there will always be a risk that the SNPs selected with most power (e.g. highest F_{ST} values), will be biased towards the most differentiated populations and will not allow for assignment to more finely differentiated groups. When dealing with large numbers of SNP markers, automated methods for selecting loci with the most power across a range of application scenarios are required; simply ranking SNPs by F_{ST} values is unlikely to lead to an optimum, minimal panel of markers for complex assignment problems, as it is particular combinations of loci that are likely to contain the highest discrimination power.

Three different approaches for locus selection have been developed together with accompanying software. WHICHLOCI (Banks *et al.* 2003) initially estimates the assignment power of individual loci from empirical data and ranks them according to individual assignment (and/or misassignments). In a second round of assignment, loci are added to an assignment trial from the top of the individual power list until the user specified level of accuracy is achieved. The programme and approach is relatively simple and straightforward. However, an important caveat is that the programme does not explore the potential power of certain combinations of loci, which may maximize IA, but may not include loci from the top of the list. An alternative approach is genetic algorithm-based feature selection (GAFS, Topchy *et al.* 2004). This programme uses a 'genetic algorithm' optimization technique, by exploring different locus combinations where the highest classification accuracy is the parameter of interest that is being searched for. The programme works on many solutions simultaneously in contrast to other optimization algorithms using incremental improvement (see above). Although the programme allows for an exhaustive search of all potential combinations, it may not be computationally feasible to explore all combinations, thereby leaving potentially highly discriminatory combinations unexplored. The third and most recently described option is 'backward elimination locus selection' using the programme BELS (Bromaghin 2008). The programme excludes each locus in the baseline data temporarily, and the baseline accuracy for assignment (or Mixed Stock Analysis) of remaining loci is evaluated iteratively. After all loci have been evaluated, the locus causing the least power reduction is permanently excluded. The procedure is repeated until only one locus is left or the level of accuracy reaches a user-defined minimum. The advantage of the programme is that (like GAFS) it exploits possible synergistic effects among loci. The downside is that with many loci and populations, it takes a long time to run on a standard desktop computer. Another shortcoming of the BELS procedure is in cases of forensic assignment where selection for the smallest subset of loci, providing 100% correct assignment is the goal. In this case, the programme is unable to rank loci as elimination of any locus from the full data set will not lead to a drop in overall assignment power (100%). Instead, a reverse procedure where loci are added according to their individual assignment power and subsequently eliminated using subsets of loci where assignment power is below 100% could be applied (J. Bromaghin, personal communication). Overall, it appears that the two latter

programmes represent the most optimal approaches for SNP loci under selection, as they search for 'synergistic' combinations of loci providing the highest overall level of assignment power regardless of their individual power.

A final note of caution for the selection of particular loci with elevated assignment power was pointed out in a recent paper by Anderson (2010). The programmes described in this section all use the same data for ranking loci and assessing their power, leading to biased and over-optimistic estimates of assignment power. Instead, Anderson suggested a procedure called THL (training, holdout, leave-one-out), where a subset of samples (training samples) is used for selection of highly informative loci to be included in the final panel of loci. These samples are combined with another subset of data (the holdout samples) to form the baseline for assignment using the final panel. By assigning the holdout samples using the full baseline sample employing a leave-one-out procedure, it is possible to separate the process of locus selection or 'high grading' from the evaluation of assignment power, while at the same time making use of the whole data set. This approach should be encouraged and implemented as a standard for evaluation of assignment power of loci under selection.

Power for detecting population differentiation. A recent paper by Morin *et al.* (2009) addresses the issue of the number of SNPs and sample size that should be used to maximize statistical power to identify evolutionary significant units (ESUs) and demographic independents units (DIPs) using the programme POWSIM (Ryman & Palm 2006). The 'effect sizes', i.e. the magnitude of differentiation required to detect two scenarios was F_{ST} = 0.2 and F_{ST} = 0.0025. The study assessed sample sizes within 10–100, number of loci 10–75 and MAFs 0.01–0.5. Overarching results showed that approximately 30 neutral loci were required to detect ESUs (N_em = 0.1), while identification of DIPs may require >75 loci. Different MAFs had little effect on power; haplotypes (linked loci) from different SNPs within the same locus could improve power, though sample size had a strong effect on power. For example, with 75 SNPs and F_{ST} = 0.0025, an increase in sample size from 50 to 100 provided a twofold increase in power (proportion of significant tests) from 0.4 to 0.8. Accordingly, if the aim is specifically to address the issue of microgeographical population structure, it may be advisable to use relatively large sample sizes. Also, including loci suspected to be under selection may increase power to detect differentiation; however, the stability of the pattern has to be investigated because contemporary selection may alter allele frequencies even within a cohort (see Nielsen *et al.* 2009a).

Glover *et al.* (2010) compared the IA resolution between analyses with 309 mapped SNPs (global F_{ST} −0.002 to 0.316; only one 'outlier locus') and 14 microsatellite markers (global F_{ST} 0.033–0.115) in wild and domesticated strains of Atlantic salmon (*Salmo salar*). They found that proportions of correctly assigned individuals was 0.65, 0.73 and 0.73 when assigned with 14 microsatellites, 300 SNPs and 195 'mapped' (>1 cM) SNPs, respectively. Overall, assignment was best (80% correct) when ~100 unlinked SNP loci were used. Above 100 loci, assignment success decreased. Comparing marker types, the most informative 15 salmon SNPs matched the level of assignment achieved by the most informative four microsatellite loci (ranked by maximizing allelic variation). If linkage information is available, Structure (Pritchard *et al.* 2000; Falush *et al.* 2003) may outperform Geneclass (Cornuet *et al.* 1999), as Structure enables the use of a linkage model, taking marker distance into consideration in computations, whereas Geneclass treats loci as independent. In the study by Glover *et al.* (2010) using Structure, the use of a linkage model led to 88% correct self-assignment when using 300 SNPs, whereas correct assignment was 80% with Geneclass. This study suggests that the identification of a highly informative set of SNPs from a larger panel is likely to give significantly more accurate individual genetic self-assignment compared to any combination of microsatellite loci. However, there is a risk of an upwards bias of the estimates of assignment success when 'high-grading' loci, as described by Anderson (2010). The study by Glover *et al.* (2010) also underlines the importance of using an appropriate method for modelling the statistical power and assignment resolution when choosing subsets of markers for targeted assignment analyses.

Conclusions

In several of the aforementioned sections, attention has been drawn to some of the concerns associated with the discovery of SNPs from NGS data. Some of these issues, such as the bias in genome coverage achieved, or the complications of not having a reference genome, are being dealt with by advances in technology (e.g. reducing the bias in terminal end sequencing (Korbel *et al.* 2007), paired-end reads for sequence assembly without a reference sequence (Li *et al.* 2010), also see Harismendy *et al.* 2009 for an evaluation of the different issues between platforms and Everett *et al.* 2011 for an assessment of the potential to assemble sequences to publicly available EST databases). Other major drawbacks such as the conversion rate from NGS data to validated SNPs, and the inherent ascertainment bias in the data still need practical solutions (for reviews see Hudson 2008; Shendure & Ji

2009; Garvin *et al.* 2010). NGS is one of the most powerful tools currently available, but its use must be undertaken with its limitations in mind. Meanwhile major advances in sequencing—such as the third generation technologies—are promising to resolve many of the difficulties with the current systems with less expensive, longer read, more accurate systems promised in the near future (Eid *et al.* 2009; Metzker 2009; Rusk 2009). However, although it has been suggested that ecologists may soon be able to perform population genetics at a genome, rather than a gene level (Hudson 2008), these technologies are likely to remain out of reach for the majority of studies on nonmodel organisms for the foreseeable future. Additionally, the replacement of SNP genotyping by the analysis of the full genome sequence data is also currently out of reach for the majority of nonmodel species.

The continued increase in speed and decrease in cost for SNP genotyping nonmodel organisms is undoubtedly going to lead to further major changes in relation to the availability of data on a genomic scale for population genetic analysis in the near future. Currently, we are in a transition period where population structure is typically inferred from relatively few genetic markers for some wild organisms, while thousands of markers and even whole genomes (Hohenlohe *et al.* 2010) are being analysed in others. Accordingly, we expect to see an increased movement towards genome wide analyses to gain a general understanding of the relative importance of neutral and adaptive processes in wild populations. Such a development will result in a conceptual change as it will no longer be feasible to manually edit or check data quality. In turn, further developments will be required in relation to statistical tools and associated software for analysing data orders of magnitude larger than is currently standard, some of which have been highlighted above. However, the fundamental principles of population genetics remain the same and specific research questions will continue to require appropriate analysis dependant on the nature of the markers used.

Although the data sets that we have access to are increasing in size, there will continue to be a need for small panels of 'genetic tags' for ecological, management and forensic purposes where the assignment of individuals and groups of individuals to the population of origin is desired. We expect these applications to grow tremendously and become commonplace as the costs of genotyping decline progressively. To generate added momentum, there is an enhanced need for genomic data for nonmodel taxa, from where the high grading of the most informative loci for individual assignment can take place to create cost-effective panels of minimum size with maximum power.

Acknowledgements

The discussions that formed the basis of this manuscript began at a FishPopTrace Consortium workshop held in February 2010. We thank all the members of the consortium present at that meeting, especially L. Bargelloni, for their contributions to the discussions; we also thank the two anonymous reviewers for their helpful comments.

Conflict of interest

The authors have no conflict of interest to declare and note that the sponsors of the issue had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805–1814.

Akey JM, Zhang K, Xiong M, Jin L (2003) The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Molecular Biology and Evolution*, **20**, 232–242.

Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*, **24**, 1–20.

Allendorf FW, Luikart G (2007) *Conservation and the Genetics of Populations*. Blackwell Publishing, Oxford, UK.

Anderson EC (2010) Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources*, **10**, 701–710.

Anderson EC, Garza JC (2006) The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, **172**, 2567–2582.

Anderson TJC, Nair S, Sudimack D *et al.* (2005) Geographical distribution of selected and putatively neutral SNPs in Southeast Asian malaria parasites. *Molecular Biology and Evolution*, **22**, 2362–2374.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics*, **9**, 323.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Ball AD, Stapley J, Dawson DA, Birkhead TR, Terry Burke T, Slate J (2010) A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (Taeniopygia guttata). *BMC Genomics*, **11**, 218.

Balzer S, Malde K, Lanzén A, Sharma A, Jonassen I (2010) Characteristics of 454 pyrosequencing data – enabling realistic simulation with flowsim. *Bioinformatics*, **26**, i420–i425.

Banks MA, Eichert W, Olsen JB (2003) Which genetic loci have greater population assignment power? *Bioinformatics*, **19**, 1436–1438.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. *Nature Genetics*, **40**, 340–345.

Beaumont MA (2005) Adaptation and speciation: what can F_{ST} tell us? *Trends in Ecology and Evolution*, **20**, 435–440.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London – Series B: Biological Sciences*, **263**, 1619–1626.

Bonin A (2008) Population genomics: a new generation of genome scans to bridge the gap with functional genomics. *Molecular Ecology*, **17**, 3583–3584.

Bonin A, Taberlet P, Miaud C, Pompanon F (2006) Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Molecular Biology and Evolution*, **23**, 773–783.

Brieuc M, Naish K (2011) Detecting signatures of positive selection in partial sequences generated on a large scale: pitfalls, procedures and resources. *Molecular Ecology Resources*, **11** (Suppl. 1), 172–183.

Bromaghin J (2008) BELS: backward elimination locus selection for studies of mixture composition or individual assignment. *Molecular Ecology Resources*, **8**, 568–571.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) Single nucleotide polymorphisms (SNPs) as markers in phylogeography. *Trends in Ecology & Evolution*, **18**, 249–256.

Cai JJ, Macpherson JM, Sella G, Petrov DA (2009) Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genetics*, **5**, e1000336.

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. *Nature*, **429**, 446–452.

Castaño-Sánchez C, Smith TPL, Wiedmann RT *et al.* (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 559.

Chen C, Durand E, Forbes F, Francois O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, **7**, 747–756.

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. *Genome Research*, **20**, 393–402.

Coop G, Pickrell JK, Novembre J *et al.* (2009) The role of geography in human adaptation. *PLoS Genetics*, **5**, e1000500.

Corander J, Siren J, Arjas E (2008) Bayesian spatial modeling of genetic population structure. *Computational Statistics*, **23**, 111–129.

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, **153**, 1989–2000.

Davidson EH, Rast JP, Oliveri P *et al.* (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.

De Carvalho D, Ingvarsson PK, Joseph J *et al.* (2010) Admixture facilitates adaptation from standing variation in the European aspen (*Populus tremula* L.), a widespread forest tree. *Molecular Ecology*, **19**, 1638–1650.

Delaneau O, Coulonges C, Zagury JF (2008) Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, **9**, 540.

Eid J, Fehr A, Gray J *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Eveno E, Collada C, Guevara MA *et al.* (2008) Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. *Molecular Biology and Evolution*, **25**, 417–437.

Everett M, Grau E, Seeb J (2011) Short reads and non-model species: exploring the complexities of next generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources*, **11** (Suppl. 1), 93–108.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Foll M, Gaggiotti O (2006) Identifying the environmental factors that determine the genetic structure of Populations. *Genetics*, **174**, 875–891.

Foll M, Gaggiotti O (2008) A Genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Freamo H, O'Reilly P, Berg P, Lien S, Boulding E (2011) Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. *Molecular Ecology Resources*, **11** (Suppl. 1), 243–256.

Gaggiotti OE, Bekkevold D, Jorgensen HBH *et al.* (2009) Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: atlantic herring as a case study. *Evolution*, **63**, 2939–2951.

Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*, **10**, 915–934. (doi:10.1111/j.1755-0998.2010.02891.x).

Glover KA, Hansen MM, Lien S, Als TD, Høyheim B, Skaala Ø (2010) A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genetics*, **11**, 2–12.

Goudet J (1995) fstat version 1.2: a computer program to calculate *F*-statistics. *Journal of Heredity*, **86**, 485–486.

Goudet J (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, **5**, 184–186.

Guillot G, Foll M (2009) Correcting for ascertainment bias in the inference of population structure. *Bioinformatics*, **25**, 552–554.

Guillot G, Santos F (2009) A computer program to simulate multilocus genotype data with spatially auto-correlated allele frequencies. *Molecular Ecology Resources*, **9**, 1112–1120.

Guillot G, Mortier F, Estoup A (2005) GENELAND: a computer package for landscape genetics. *Molecular Ecology Notes*, **5**, 712–715.

Guillot G, Leblois R, Coulon A, Frantz AC (2009) Statistical methods in spatial genetics. *Molecular Ecology*, **18**, 4734–4756.

Harismendy O, Ng PC, Strausberg RL *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, **10**, R32.

Hauser L, Baird M, Hilborn R, Seeb L, Seeb J (2011) An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. *Molecular Ecology Resources*, **11** (Suppl. 1), 150–161.

Hayes B, Laerdahl J, Lien S *et al.* (2007) An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture*, **265**, 82–90.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution*, **59**, 1633–1638.

Hemmer-Hansen J, Nielsen E, Meldrup D, Mittelholzer C (2011) Identification of single nucleotide polymorphisms in candidate genes for growth and reproduction in a nonmodel organism; the Atlantic cod, Gadus morhua. *Molecular Ecology Resources*, **11** (Suppl. 1), 71–80.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresk WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLOS Genetics*, **6**, e1000862.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics*, **10**, 639–650.

Hudson M (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Joost S, Kalbermatten M, Bonin A (2008) Spatial analysis method (SAM): a software tool combining molecular and environmental data to identify candidate loci for selection. *Molecular Ecology Resources*, **8**, 957–960.

Jost L (2008) G_{ST} and its relatives do not measure differentiation. *Molecular Ecology*, **17**, 4015–4026.

Kaeuffer R, Reale D, Coltman DW, Pontier D (2007) Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity*, **99**, 374–380.

Kalinowski ST (2002) How many alleles per locus should be used to estimate genetic distances? *Heredity*, **88**, 62–65.

Karlsson S, Moen T, Lien S, Glover K, Hindar K (2011) Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP-chip. *Molecular Ecology Resources*, **11** (Suppl. 1), 236–242.

Kauer M, Dieringer D, Schlotterer C (2003) Nonneutral admixture of immigrant genotypes in African *Drosophila melanogaster* populations from Zimbabwe. *Molecular Biology and Evolution*, **20**, 1329–1337.

Kerstens HH, Crooijmans RP, Veenendaal A *et al.* (2009) Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics*, **10**, 479–489.

Korbel JO, Urban AE, Affourtit JP *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.

Krawczak M (1999) Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis*, **20**, 1676–1681.

Lao O, van Duijn K, Kersbergen P, de Knijff P, Kayser M (2006) Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *American Journal of Human Genetics*, **78**, 680–690.

Laval G, Patin E, Barreiro LB, Quintana-Murci L (2010) Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS ONE*, **5**, e10284.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Li R, Fan W, Tian G *et al.* (2010) The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317.

Liu N, Chen L, Wang S, Oh C, Zhao H (2005) Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics*, **6**(Suppl. 1), S26.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology and Evolution*, **20**, 136–142.

Marshall JM, Weiss RE (2006) A Bayesian heterogeneous analysis of variance approach to inferring recent selective sweeps. *Genetics*, **173**, 2357–2370.

Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations GENETICS, **166**(1): 351–372.

Maynard Smith J, Haigh J (1974) Hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.

Metzker M (2009) Sequencing in real time. *Nature Biotechnology*, **27**, 150–151.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Moen T, Hayes B, Nilsen F *et al.* (2008) Identification and characterisation of novel SNP markers in Atlantic cod: evidence for directional selection. *BMC Genetics*, **9**, 18.

Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theoretical and Applied Genetics*, **120**, 1525–1534.

Morcillo-Suarez C, Alegre J, Sangros R *et al.* (2008) SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data. *Bioinformatics*, **24**, 1643–1644.

Morin PA, McCarthy M (2007) Highly accurate SNP genotyping from historical and low-quality samples. *Molecular Ecology Notes*, **7**, 937–946.

Morin PA, Luikart G, Wayne RK, SNP_Workshop_Group (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, **19**, 208–216.

Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, **9**, 66–73.

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology*, **17**, 3599–3613.

Nicholson G, Smith AV, Jónsson F *et al.* (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 695–715.

Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**, 931–942.

Nielsen R, Signorovitch J (2003) Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical Population Biology*, **63**, 245–255.

Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics*, **168**, 2373–2382.

Nielsen EE, Hansen MM, Meldrup D (2006) Evidence of microsatellite hitch-hiking selection in Atlantic cod (*Gadus morhua* L.): implications for inferring population structure in non-model organisms. *Molecular Ecology*, **15**, 3219–3229.

Nielsen EE, Hemmer-Hansen J, Poulsen NA *et al.* (2009a) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology*, **9**, 276.

Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009b) Population genomics of marine fishes: identifying adaptive variation in space and time. *Molecular Ecology*, **18**, 3128–3150.

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646–649.

Palumbi SR, Baker CS (1994) Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. *Molecular Biology and Evolution*, **11**, 426–435.

Paschou P, Ziv E, Burchard EG *et al.* (2007) PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics*, **3**, 1672–1686.

Patterson N, Price A, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, e190.

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A (2004) GeneClass2: a Software for Genetic Assignment and First-Generation Migrant Detection. *Journal of Heredity*, **95**, 536–539.

Polanski A, Kimmel M (2003) New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, **165**(1): 427–436.

Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, **155**, 945–959.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, **81**, 559–575.

Reeves PA, Richards CM (2009) Accurate Inference of Subtle Population Structure (and Other Genetic Discontinuities) Using Principal Coordinates. *PLoS ONE*, **4**, e4269.

Robertson A (1975) Remarks on the Lewontin-Krakauer test. *Genetics*, **80**, 396–396.

Romero IG, Manica A, Goudet J, Handley LL, Balloux F (2009) How accurate is the current picture of human genetic variation? *Heredity*, **102**, 120–126.

Rosenberg N, Li L, Ward R, Pritchard J (2003) Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics*, **73**, 1402–1422.

Rosenblum EB, Novembre J (2007) Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. *Journal of Heredity*, **98**, 331–336.

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Molecular Ecology Resources. **8**, 103–106.

Rusk N (2009) Cheap third-generation sequencing. *Nature Methods*, **6**, 244–245.

Ryman N, Leimar O (2008) Effect of mutation on genetic differentiation among nonequilibrium populations. *Evolution*, **62**, 2250–2259.

Ryman N, Palm S (2006) POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Molecular Ecology Notes*, **6**, 600–602.

Sabeti PC, Varilly P, Fry B *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–919.

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, **78**, 629–644.

Schlötterer C (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics*, **160**, 753–763.

Schlötterer C, Harr B (2002) Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in *Drosophila melanogaster*. *Molecular Ecology*, **11**, 947–950.

Shendure J, Ji H (2009) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.

Smith CT, Antonovich A, Templin WD, Elfstrom CM, Narum SR, Seeb LW (2007) Impacts of marker class bias relative to locus-specific variability on population inferences in Chinook salmon: a comparison of single-nucleotide polymorphisms with short tandem repeats and allozymes. *Transactions of the American Fisheries Society*, **136**, 1674–1687.

Smith M, Pascal C, Grauvogel Z, Habicht C, Seeb J, Seeb L (2011) Multiplex preamplification PCR and microsatellite validation allows accurate single nucleotide polymorphism (SNP) genotyping of historical fish scales. *Molecular Ecology Resources*, **11** (Suppl. 1), 257–266.

Stephens M, Smith N, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

Storz JF, Kelly JK (2008) Effects of spatially varying selection on nucleotide diversity and linkage disequilibrium: insights from deer mouse globin genes. *Genetics*, **180**, 367–379.

Topchy A, Scibner K, Punch W (2004) Accuracy-driven loci selection and assignment of individuals. *Molecular Ecology Notes*, **4**, 798–800.

Vasemägi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology*, **14**, 3623–3642.

Vitalis R, Dawson K, Boursot P (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics*, **158**, 1811–1823.

Vitalis R, Dawson K, Boursot P, Belkhir K (2003) DetSel 1.0: a computer program to detect markers responding to selection. *Journal of Heredity*, **94**, 429–431.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A Map of Recent Positive Selection in the Human Genome. *PLoS Biology*, **4**, e72.

Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K (2001) The discovery of single-nucleotide polymorphisms – and inferences about human demographic history. *American Journal of Human Genetics*, **69**, 1332–1347.

Waples R (2010). Perspective. High grading bias: subtle problems with assessing power of selected subsets of loci for population assignment. *Molecular Ecology*, **19**, 2599–2601.

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

White C, Selkoe KA, Watson J, Siegel DA, Zacherl DC, Toonen RJ (2010) Ocean currents help explain population genetic structure. *Proceedings of the Royal Society B-Biological Sciences*, **277**, 1685–1694.

Woolfe A, Goodson M, Goode DK *et al.* (2005) Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biology*, **3**, e7. doi:10.1371/journal.pbio.0030007.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Xu H, Sarkar B, George V (2009) A new measure of population structure using multiple single nucleotide polymorphisms and its relationship with FST. *BMC Research Notes*, **2**, 21.

Xu S, Gupta S, Jin L (2010) PEAS V1.0: a package for elementary analysis of SNP data. *Molecular Ecology Resources*, **10**, 1085–1088. Online Early.

Chapter 6

SNP discovery using next generation transcriptomic sequencing in Atlantic Herring (*Clupea harengus*)

Published in *PLoS ONE*

OPEN ACCESS Freely available online

PLOS ONE

SNP Discovery Using Next Generation Transcriptomic Sequencing in Atlantic Herring (*Clupea harengus*)

Sarah J. Helyar[1,2*☯], Morten T. Limborg[3☯], Dorte Bekkevold[3], Massimiliano Babbucci[4], Jeroen van Houdt[5], Gregory E. Maes[5], Luca Bargelloni[4], Rasmus O. Nielsen[6], Martin I. Taylor[1], Rob Ogden[7], Alessia Cariani[8], Gary R. Carvalho[1], FishPopTrace Consortium[1], Frank Panitz[6]

1 Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, College of Natural Sciences, Bangor University, Bangor, Gwynedd, United Kingdom, **2** Food Safety, Environment & Genetics, Matís, Reykjavík, Iceland, **3** National Institute of Aquatic Resources, Technical University of Denmark, Silkeborg, Denmark, **4** Department of Comparative Biomedicine and Food Science, University of Padova, Legnaro, Italy, **5** Laboratory of Biodiversity and Evolutionary Genomics, Katholieke Universiteit Leuven, Leuven, Belgium, **6** Department of Molecular Biology and Genetics, Faculty of Science and Technology, Aarhus University, Tjele, Denmark, **7** TRACE Wildlife Forensics Network, Royal Zoological Society of Scotland, Edinburgh, United Kingdom, **8** Department of Experimental and Evolutionary Biology, University of Bologna, Bologna, Italy

Abstract

The introduction of Next Generation Sequencing (NGS) has revolutionised population genetics, providing studies of non-model species with unprecedented genomic coverage, allowing evolutionary biologists to address questions previously far beyond the reach of available resources. Furthermore, the simple mutation model of Single Nucleotide Polymorphisms (SNPs) permits cost-effective high-throughput genotyping in thousands of individuals simultaneously. Genomic resources are scarce for the Atlantic herring (*Clupea harengus*), a small pelagic species that sustains high revenue fisheries. This paper details the development of 578 SNPs using a combined NGS and high-throughput genotyping approach. Eight individuals covering the species distribution in the eastern Atlantic were bar-coded and multiplexed into a single cDNA library and sequenced using the 454 GS FLX platform. SNP discovery was performed by *de novo* sequence clustering and contig assembly, followed by the mapping of reads against consensus contig sequences. Selection of candidate SNPs for genotyping was conducted using an *in silico* approach. SNP validation and genotyping were performed simultaneously using an Illumina 1,536 GoldenGate assay. Although the conversion rate of candidate SNPs in the genotyping assay cannot be predicted in advance, this approach has the potential to maximise cost and time efficiencies by avoiding expensive and time-consuming laboratory stages of SNP validation. Additionally, the *in silico* approach leads to lower ascertainment bias in the resulting SNP panel as marker selection is based only on the ability to design primers and the predicted presence of intron-exon boundaries. Consequently SNPs with a wider spectrum of minor allele frequencies (MAFs) will be genotyped in the final panel. The genomic resources presented here represent a valuable multi-purpose resource for developing informative marker panels for population discrimination, microarray development and for population genomic studies in the wild.

Citation: Helyar SJ, Limborg MT, Bekkevold D, Babbucci M, van Houdt J, et al. (2012) SNP Discovery Using Next Generation Transcriptomic Sequencing in Atlantic Herring (*Clupea harengus*). PLoS ONE 7(8): e42089. doi:10.1371/journal.pone.0042089

Editor: Arnar Palsson, University of Iceland, Iceland

Received March 7, 2012; **Accepted** July 2, 2012; **Published** August 7, 2012

Funding: The research leading to these results received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no KBBE-212399 (FishPopTrace). In addition, MTL received financial support from the European Commission through the FP6 projects UNCOVER (Contract No. 022717) and RECLAIM (Contract No. 044133). GM is a post-doctoral researcher funded by the Scientific Research Fund Flanders (FWO-Flanders). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sarah.helyar@matis.is

☯ These authors contributed equally to this work.

Introduction

Population genomic approaches have been revolutionised by Next Generation Sequencing (NGS) technologies such as 454 (Roche) and Illumina sequencing. These developments facilitate genome-wide analyses of genetic variation across populations of non-model organisms [1,2], allowing a range of evolutionary questions to be investigated effectively for the first time. Marine fishes are excellent model systems for studying adaptation due to their large geographic ranges that frequently encompass strong environmental gradients and their large population sizes that increase the relative strength of selection over drift [3]. Moreover, many marine fishes are under extreme anthropogenic pressure and there is an urgent need for genomic tools to identify population structure and boundaries to allow effective management [4]. Additionally the forensic identification of fish and fish products throughout the food processing chain from net to plate would assist in the fight against Illegal, Unreported and Unregulated (IUU) fishing, currently a priority for the European Union [5] and globally [6]. SNPs are the optimal marker for this type of application, but large SNP panels are currently available for few marine fish species (e.g. Atlantic cod (*Gadus morhua*) [7]; European hake (*Merluccius merluccius*) [8]). Thus, the development of genomic resources for marine fish is urgently required for evolutionary, conservation and management perspectives.

The strategy used for SNP development in non-model organisms is dependent on the availability of genomic information from closely related species. If such resources are available, PCR amplicons (homologous to regions in the reference genome) can be sequenced and SNPs identified (however, these are intrinsically limited in the number of SNPs that can be identified). Without a reference genome, three principal strategies for genome-wide SNP discovery can be applied; whole genome sequencing and assembly, genome complexity reduction and sequencing methods (e.g. RRL and RAD-seq) and cDNA sequencing (RNA-seq). While whole genome sequencing has now been completed for species with large complex genomes (for example: panda (*Ailuropoda melanoleura*) [9]; cacao (*Theobroma cacao*) [10]), this remains outside the scope of most studies, as in general the *de novo* assembly of larger, repeat-rich or polyploid genomes requires additional information (e.g. physical BAC maps or paired-end libraries) and extensive bioinformatic capacity in order to build the large, computationally intensive, structured sequence scaffolds [11]. Genomic libraries which sequence a small fraction of the genome (typically 3–5%) require a high level of coverage for contig assembly and detection of SNP variants (see [12–14] for applications). Deep sequencing of cDNA libraries provides an attractive approach to achieve the high sequence coverage needed for *de novo* contig assembly and SNP prediction, as only a small percentage of the genome is accounted for by the transcriptome. Another advantage of transcriptome sequencing is the information produced concerning functional genetic variation in specific genes which may be under selection; these can then be targeted to evaluate gene expression profiles. The ability to examine both neutral variation and genomic regions under selection provides researchers with unprecedented tools for understanding local adaptation of wild populations at the molecular level.

Atlantic herring (*Clupea harengus*) is an abundant and ecologically highly diverse species, occurring with a more or less continuous distribution in the North-Atlantic benthopelagic zone. Habitats are distributed across highly diverse environments, from temperate (33°N) to arctic (80°N) and at salinities from oceanic (~35 ppt) to brackish (down to 3 ppt). In spite of its large ecological range, studies using "neutral" microsatellites have unanimously reported weak population differentiation that is statistically significant only on regional scales [15–17]. However, despite relatively high levels of gene flow among populations, evidence of local adaptation has been identified in the Atlantic herring in the Baltic Sea using microsatellite loci [18,19]. Therefore it is expected that analyses with transcriptome-wide coverage applying hundreds of markers associated with adaptive and neutral variation will provide novel insights into the role of selective and demographic processes in shaping population structure.

We describe transcriptome-based SNP development in Atlantic herring using a Roche 454 GS FLX (hereafter 454) sequencing approach. Our aim was three-fold; 1) to develop a SNP assay exhibiting minimal ascertainment bias across east Atlantic populations, 2) to test the applicability of *in silico* SNP detection utilizing a combined SNP screening and validation approach as a cost efficient way of obtaining population genomic resources, and 3) to establish a transcriptome resource for tissue-specific gene expression profiling and microarray development. We present, to our knowledge, one of the first studies describing SNP discovery in a non-model marine fish based on transcriptome sequencing using NGS.

Materials and Methods

cDNA Library Construction and 454 Sequencing

SNP development was based on muscle samples from eight fish collected from four locations from across the eastern Atlantic (Figure 1). These locations were chosen to maximise geographic coverage and environmental differences, thereby minimising potential ascertainment bias. Approximately 5g of muscle tissue was taken from each of two individuals (male and female) from each location and immediately placed in RNAlater (Invitrogen) and after 12 hours at 4°C, were stored at −80°C. Total RNA was extracted using the RNeasy Lipid Tissue Mini Kit (Qiagen). The Oligotex mRNA Mini Kit (Qiagen) was used to isolate mRNA, and non-normalised cDNA was synthesized using the SuperScript Double-stranded cDNA Synthesis Kit (Invitrogen). A multiplex sequencing library was prepared by pooling equal amounts of cDNA from all eight individuals, where two specific 10-mer barcoding oligonucleotides were ligated to each individual sample to allow post-sequencing identification of sequences (modified from [20]). High-throughput sequencing was performed on a 454 sequencer according to the manufacturers' protocol.

Sequence Processing and Assembly

Sequences were first de-multiplexed using the barcoding tags (sfffile tool, Roche 454 analysis software) and sorted by sample. Mitochondrial sequences were removed from the data set by mapping the reads against the Atlantic herring mitochondrial genome (Genbank accession NC_009577 [21]) using the Roche 454 gsMapper software. RepeatMasker [22] was used to identify and mask repetitive and low complexity regions within the reads by using the zebrafish (*Danio rerio*) repeat library. Reads were cleaned for short sequences (<50 bp) and low quality regions using SeqClean (http://compbio.dfci.harvard.edu/tgi/software/). Sequence clustering was performed in two steps; initial clustering was performed using CLC Genomics Workbench (CLCbio, Denmark), the resulting ace file sequences were then assembled 'per contig' in CAP3 [23]. The consensus sequences for the contigs produced by this assembly were then used as a reference for mapping reads in the subsequent *in silico* SNP detection.

SNP Detection

To identify candidate SNPs, all contig specific reads from the CAP3.ace files were re-mapped onto the consensus sequence and candidate SNPs were identified using GigaBayes [24]. This program scans each position of the assembly for the presence of at least two SNP alleles and calculates the probability of a given site being polymorphic using a Bayesian approach. No insertion or deletion variants (InDels) were considered and the polymorphism rate was set to 0.003. A minimum contig depth of four reads covering the polymorphic site and a minimum of two reads for the rare allele were required for a site to be considered as a putative SNP. All contigs containing SNPs were filtered to remove instances in which the alternative allele of the SNP was only identified in a single individual, as these may either represent false positives or may lead to strong ascertainment bias.

Microsatellite Sequence Screening

Microsatellites are an important resource for smaller scale studies in population genetics, microsatellites within expressed genomic regions have been shown to produce clearer genotyping results as there are fewer null alleles and stutter bands [25,26]; therefore the contig library developed here was screened to detect repeat regions. Assembled contigs were screened for microsatellite repeats using MsatCommander [27] a Python program which

Figure 1. Location of the 18 samples used in this study. The eight sequenced ascertainment individuals (2 per location) came from the four sampling sites denoted in red.
doi:10.1371/journal.pone.0042089.g001

locates microsatellite repeats (di-, tri-, tetra-, penta-, and hexa-nucleotide repeats) within fasta-formatted sequences or consensus files. MsatCommander then uses Primer3 [28] to screen sequences containing microsatellite loci for high-quality PCR primer sites within the flanking regions for 'potentially amplified loci' (PALs [29]).

Contig Annotation

Contigs were annotated using the Basic Local Alignment Search Tool (BLAST) against multiple sequence databases. Blastn searches (E-value cut-off $<1.0\ E^{-5}$) were conducted against all annotated transcripts of *Gasterosteus aculeatus*, *Tetraodon nigroviridis*, *Oryzias latipes*, *Takifugu rubripes*, *Danio rerio* and *Homo sapiens* available through the Ensembl Genome Browser, and against all unique transcripts for *D. rerio*, *H. sapiens*, *O. latipes*, *T. rubripes*, *Salmo salar*, and *Oncorhynchus mykiss* in the NCBI UniGene database. Blastx searches were conducted (E-value cut off $<1.0\ E^{-3}$) against the UniProtKB/SwissProt and UniProtKB/TrEMBL databases. Lastly Blastx searches were performed against all annotated proteins from the transcriptomes of *G. aculeatus*, *T. nigroviridis*, *O. latipes*, *T. rubripes*, *D. rerio* and *H. sapiens* available through the Ensembl Genome Browser.

To predict the effect of the mutation underlying each SNP at the amino acid level, a pipeline was developed to predict the reading frame for each SNP-containing contig. All contigs containing SNPs were first blasted against six peptide sequence databases (Ensembl genome assembly for *G. aculeatus*, *T. nigroviridis*, *O. latipes*, *T. rubripes*, *D. rerio* and the Swissprot database) using the Blastx function (E-value cut-off $<1.0\ E^{-3}$). For each SNP containing contig the best match was selected and the aligned sections of the query were saved. Subsequently, two 121 bp sequences per SNP (i.e. 60 bp up/down-stream of the SNP position, one sequence for each allele) were produced, these were used in a Blastx analysis against the file retrieved from the peptide sequences (E-value cut-off $<1.0\ E^{-10}$), and were then compared to determine if the SNP represented a synonymous or non-synonymous mutation.

Selection of Candidate SNPs for Genotyping Assay

SNPs were validated following an *in silico* protocol, aimed at minimising validation costs, whilst also minimising subsequent locus dropout. SNP selection was based on the results from the Illumina Assay Design Tool, detection of putative intron-exon boundaries within the flanking regions of candidate SNPs, and a visual evaluation of the quality of contig sequence alignments. The SNPScore from the Illumina Assay Design Tool (referred to as the Assay Design Score/ADS) utilises factors including template GC content, melting temperature, sequence uniqueness, and self-complementarity to filter the candidate SNPs prior to further inspection. The Assay Design Score (assigned between 0 and 1) is

indicative of the ability to design suitable oligos within the 60 bp up/down-stream flanking region, and the expected success of the assay when genotyped with the Illumina GoldenGate chemistry. Following the Illumina guidelines, all SNPs with a score below 0.4 were discarded; SNPs with a score above 0.4 were accepted, with SNPs scoring above 0.7 being used preferentially.

The prediction of intron-exon boundaries within the SNP flanking regions (60 bp up/down-stream of SNP position) was performed using two approaches. The first directly compared SNP-containing contigs against five high quality reference genomes for model fish species (Ensembl genome assembly for *G. aculeatus*, *T. nigroviridis*, *O. latipes*, *T. rubripes* and *D. rerio;* see Figure S1, left pipeline), using the Blastn option (E-value cut-off 10^{-5}). Blast results were then parsed via a custom Perl script considering alignment length, start and end point of the alignment to determine the best positive match (further details of the Perl script and workflow are available from the authors on request). If the 60 bp on both sides of the SNP were present in the alignment, the candidate SNP was considered to be contained within a single exon; otherwise an intron-exon boundary was assumed to be present within the 121 bp assay design region. SNPs were then assigned to one of three categories either having, or not having an intron-exon boundary predicted within the flanking region, or as not returning a significant match against any of the five blasted fish genomes. In the other approach, the likelihood of a positive match and the reliability of intron-exon boundary predictions were increased, with SNP-containing contigs used as a query in a Blast search (blastn, E-value cut-off 10^{-5}) against the corresponding transcriptome of the same five reference databases (see above). If the blast search produced a positive result, the matching transcript was downloaded from the Ensembl database, and blasted against its own genome sequence (see Figure S1, right pipeline). Within the downloaded sequence, the nucleotide position corresponding to the candidate SNP in the Atlantic herring sequence was identified based on the start and end positions of the alignment between the original contig and the Ensembl transcript. Using the projected SNP position, the flanking regions were again classified as being located on a single exon, disrupted by an intron, or not having a significant match. Results from the two approaches were compared to obtain a consensus estimate for the likelihood of an intron-exon boundary occurring within the 121 bp assay for each of the candidate SNPs.

Finally, the remaining candidate SNP contigs were visually evaluated using clview (clview; http://compbio.dfci.harvard.edu/tgi/software/) in order to rank putative SNPs within and among contigs. This was assessed by considering the overall quality of the assembly, the depth and length of alignments, and the number of mismatch sites flanking the SNP. This step was included to increase the likelihood of excluding incorrectly identified SNPs (for example; regions with alternative splicing or erroneous clustering of paralogous sequences). Within each contig, one or two SNPs receiving the highest quality score were considered for further validation (see below).

SNP Validation

Following the pipeline described above, 1,536 high scoring candidate markers were chosen for validation by high throughput genotyping assay. DNA was extracted from fin clips for 626 fish sampled from eighteen sites across the species range in the eastern Atlantic, including twenty fish from each of the four SNP discovery populations (Figure 1). The quality and quantity of DNA was checked using a Nanodrop spectrophotometer, and all samples were standardised to 70 ng/μL. Genotyping was performed using the Illumina Golden Gate platform [30], and was visualised using Illumina's GenomeStudio data analysis software (1.0.2.20706, Illumina Inc.). Only SNP assays showing clear genotype clustering, and individual samples with a call rate above 0.8 were considered for further analysis.

Cross-species Amplification

To assess the utility of developed markers in related species, two species identified from a consensus phylogeny [31], the sister species; Pacific herring (*C. pallasii*) and a more distantly related species; anchovy (*Engraulis encrasicolus)* were genotyped for the full 1,536 SNP panel.

Statistical Analyses

To assess the predictive value and utility of the different parameters used in the *in silico* SNP validation pipeline, a binomial logistic regression analysis was conducted. Two categorical variables (*Conversion* and *Polymorphism*) were evaluated which describe the outcome of the SNP assay validation; these are expected to depend on a range of candidate predictor variables (see below). *Conversion* was scored by assigning all 1,536 genotyped SNP assays as either failed (score = 0) or successfully amplified and clustered (score = 1). *Polymorphism* assigned all the successfully amplified SNP assays into monomorphic (0) or polymorphic (1) categories. Nine variables were then assessed for their predictive value in determining SNP assay conversion and polymorphism: i) number of ascertainment panel individuals supplying sequence reads at the SNP position, ii) number of sequences aligned under SNP position, iii) number of sequences with the minor allele, iv) frequency of sequences with minor allele, v) number of ascertainment individuals with the minor allele, vi) Illumina Assay Design Score (ADS), vii) outcome of the intron-exon boundary pipeline (scored as SNP assay being within a single exon, interrupted by an intron or as having no BLAST match), viii) number of reference species supporting findings from the intron-exon pipeline, and ix) neighbourhood sequence quality (determined by the number of mismatches in the flanking region alignment). To statistically test the predictive effect of the above variables for both *Conversion* and *Polymorphism* a two-step binomial logistic regression analysis was used as implemented in SPSS v12.0. All variables were included in the initial model, and a backward stepwise deletion approach was used for optimisation, in which the least informative variable is removed sequentially until only significantly contributing variables remain. A Wald χ^2 statistic was used to estimate the relative contribution from each remaining parameter.

For the successful polymorphic assays global values of observed (H_O) and expected (H_E) heterozygosity were estimated for 20 individuals from each of the four ascertainment populations (Figure 1) using GenAlEx 6.4 [32]. For these same populations deviations from Hardy-Weinberg equilibrium (HWE) and evidence of linkage disequilibrium (LD) were explored using Genepop 4.0 [33]. Significance levels for HWE and LD tests were estimated using an MCMC chain of 10,000 iterations and 20 batches. *P*-values were adjusted for multiple tests by false discovery rate (FDR) correction following Benjamini & Yekutieli [34].

Lastly, ascertainment bias, resulting from the non-random exclusion of SNPs with a low Minor Allele Frequency (MAF) from the marker panel, may occur due to the small size (n = 8) of the ascertainment panel (compared to the whole population), and the limited geographical coverage (compared to the whole species range). When markers are then genotyped on a much larger sample of individuals the resulting ascertainment bias [35,36] may affect the estimation of many evolutionary and population genetic parameters [2]. To assess the magnitude of a potential bias, the distribution of MAF in the marker panel was assessed across a

large data set covering 18 locations across the Eastern Atlantic to check for an elevated non-random exclusion of SNPs with a low MAF. An un-biased SNP panel should exhibit an "L-shape" distribution of MAF categories indicating adequate representation of low MAF SNPs [37].

Results

454 Sequencing

Results for the sequencing and SNP discovery pipeline are illustrated in Figure 2. A total of 683,503 cDNA sequences were generated from the multiplexed Atlantic herring muscle library. The reads were de-multiplexed to assign reads to one of the eight sequenced individuals according to their barcoding tag. For 8% of the raw reads no barcoding tag was identified, while the remaining 629,541 raw reads (average read length: 205 bp, Figure 3B) contained the 5′ tag sequence and could be allocated to pools per sample per geographical region (Figure 3A). Geographic pools ranged from 86,731 (English Channel) to 187,554 (Barents Sea) sequences. All 454 sequence data has been submitted to the Sequence Read Archive (SRA) under the study accession number ERP001233 (http://www.ebi.ac.uk/ena/data/view/ERP001233).

Sequence Processing and Assembly

Sequence cleaning and processing identified 5.8% of the assigned reads as having a match of at least 94% identity over 60 base pairs to Atlantic herring mitochondrial sequences and these were removed from the data set. RepeatMasker masked 1.9% of the dataset using the zebrafish repeat library. The SeqClean program removed a further 3.5% of the assigned reads due to low-complexity (n = 7,885), low quality (n = 169) or being below the minimum read length of 50 bases (n = 13,010). Lastly, some reads were trimmed, yielding a total of 571,731 reads for sequence clustering and assembly. Initially reads were clustered with CLC Genomics Workbench (CLCbio, Denmark), resulting in 16,456 clusters ranging from 200–400 bp. These were then individually re-assembled with CAP3 resulting in 19,246 contigs (some clusters produced by CLC were split into two or more contigs) and 30,344 singletons of which more than 50% could be annotated (Table 1). The majority of contigs consisted of less than 30 reads and ranged between 100–500 bp (Figure 3C-D).

SNP Detection and Annotation Results

SNP discovery with GigaBayes detected 6,331 putative SNPs in 1,991 separate contigs. The primary annotation of contig sequences is summarized in Table 1 and in more detail in Table S1.

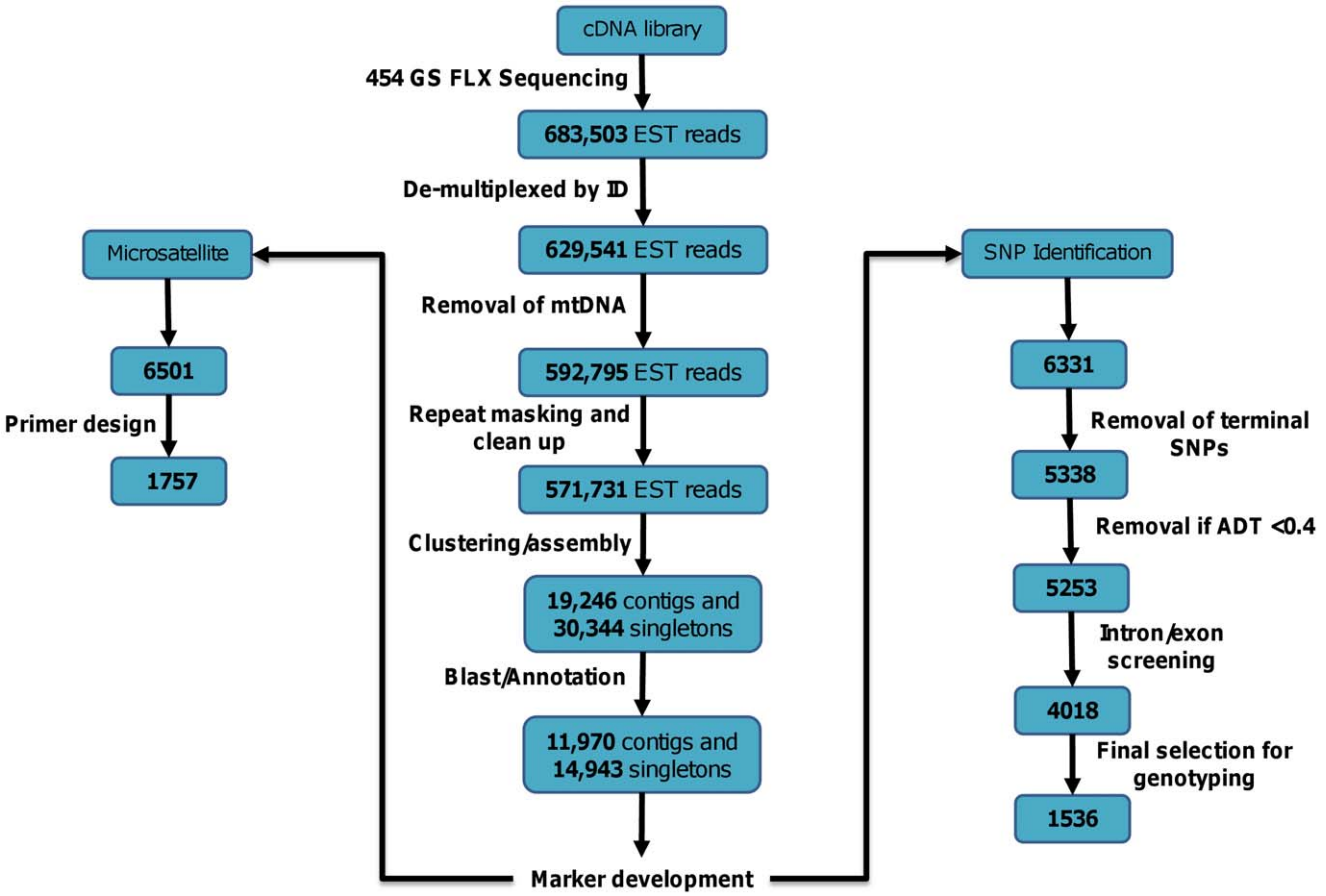


Figure 2. Schematic of transcript assembly and SNP detection pipeline. Schematic overview with numbers of reads, contigs and SNPs through the transcript assembly (centre) SNP detection (right hand side) and microsatellite detection (left hand side) pipelines (see text for more details).
doi:10.1371/journal.pone.0042089.g002

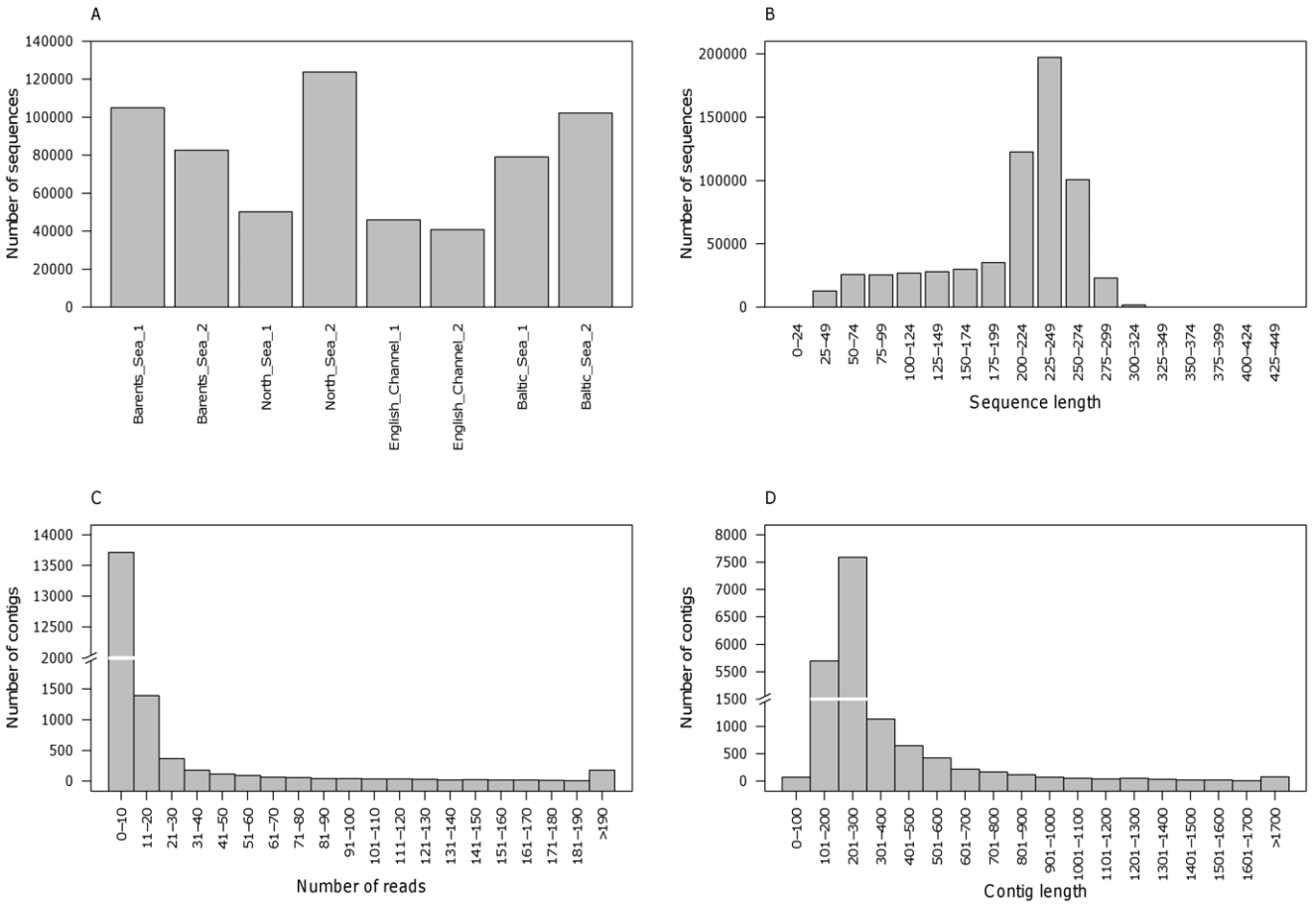


Figure 3. Summary of sequence data. A) number of sequences successfully barcoded for each of the eight ascertainment individuals; and for the combined data, B) sequence length, C) number of reads per contig and D) contig length.
doi:10.1371/journal.pone.0042089.g003

Selection of Candidate SNPs for Genotyping Assay

From the 6,331 predicted SNPs, 993 (15.6%) were located in the terminal region of the contigs and did not have the required minimum of a 60 bp flanking region to design oligos for the GoldenGate array (Figure 2). Of those remaining, 85 SNPs (1.3%) scored below the minimum value (<0.4) recommended for primer design and were not considered. 4,104 SNPs (76.8%) had high Assay Design Scores (between 0.7–1.0) and 1,149 SNPs (21.5%) had acceptable Assay Design Scores (between 0.4–0.7), all 5,253 of these were taken forward to the next stage. Of the putative SNPs screened for potential intron/exon splicing sites within the flanking regions, 1,235 (23.5%) had putative intron/exon boundaries within the flanking regions, and so were rejected. The majority (3,052, 58.1%) had no matching BLAST hits, while just 966 (18.4%) had BLAST hits which suggested that there was no intron/exon boundary present (summarised in Figure 2).

Table 1. Number of contigs and singletons obtained and successfully annotated.

	Total	Annotated	%Annotated
Contigs	19,246	11,970	62.1
Singletons	30,344	14,943	49.2
Total	49,590	26,913	54.3

doi:10.1371/journal.pone.0042089.t001

SNP Validation

From the full 1,536 panel of SNPs that were genotyped, 290 (19%) assays failed to amplify. Of the remaining 1,246 assays, 201 were monomorphic (false positives: 13%) 467 produced ambiguous clustering (30%) and 578 were polymorphic, equivalent to a conversion rate of 38%. From these 578 SNPs an open reading frame was obtained for 270 of the respective 121 bp sequences (SNP and 60 bp up/down stream), of which 66 were suggested be non-synonymous, and 204 to be synonymous, equivalent to a ratio (non-synonymous/synonymous) of 0.32 (Table S2).

Results on the predictive value of the SNP selection parameters for assay conversion (i.e. for successful amplification) show that inclusion of all of the predictor variables (see methods) marginally improves model-fitting ($\chi^2 = 18.520$, d.f. = 9, $p<0.030$). When using backward stepwise deletion of predictor variables, the Assay Design Score and number of ascertainment individuals with the minor allele were identified as the only significant predictors of assay conversion, but only the Assay Design Score showed the expected positive correlation with conversion rate (Table 2). The binomial logistic regression analysis on the polymorphic status of all successfully amplifying assays showed that when all predictor variables were included, the overall model fit was not significant ($\chi^2 = 11.554$, d.f. = 9, $p = 0.240$). However, neighbourhood se-

quence quality had a significant negative correlation with polymorphism. As before a backward stepwise deletion approach was used and this reduced the significantly contributing predictors to the number individuals in the ascertainment panel with the minor allele and the neighbourhood sequence quality which, as expected, respectively showed positive and negative correlation with SNP polymorphism (Table 3).

Estimates of H_O and H_E across the four ascertainment samples ranged from 0.00–0.63 (mean 0.18) and 0.00–0.50 (mean 0.18), respectively (Table S2). Observed heterozygosity within the four ascertainment populations revealed similar levels of diversity to the 18 sampled locations used for the SNP validation [38]. Tests for deviation from HWE for each locus and population revealed 43 out of 1,249 performed tests (3.4%) with significant deviations from HWE before correction for multiple tests. These tests were distributed among all four populations and across 35 loci. Eight tests distributed across three populations and seven loci retained significance following correction for multiple tests ($\alpha = 0.05$). Due to the presence of monomorphic loci in the four ascertainment samples, 229,094 tests for LD were performed of which 352 remained significant after correction for FDR ($\alpha = 0.05$). Of these, 14 pairs were significant in more than one of the four populations but in all cases SNPs originated from different contigs suggesting lack of close physical linkage. SNP frequency distributions of MAF categories in the full panel of 18 samples indicated little bias due to non-random selection of high frequency SNPs (Figure 4).

Cross-species Amplification and Microsatellite Detection

The majority (99%) of the 578 markers identified as polymorphic in Atlantic herring also amplified in Pacific herring, but only 12% exhibited more than one allele. Only about 10% of the 578 SNPs amplified in anchovy, and of these, only ten loci exhibited polymorphism.

MsatCommander detected 6,501 microsatellites with a repeat length of between two and seven bases with four or more repeat units in 3,741 contigs (Table 4). 27% of the microsatellites had sufficient suitable flanking sequence to enable the design of primers. Details of the microsatellites (number and type of repeat, primers, Tm and %GC) are listed in Table S3.

Discussion

This study demonstrates the *de novo* discovery of 6,331 putative SNPs based on 454 transcriptome sequencing of eight individuals covering the Northeast Atlantic distribution of the Atlantic herring. Of particular interest in the approach is the single validation and genotyping step, disposing with the traditional step of testing each SNP for amplification prior to large scale genotyping (e.g. [39,40]). The data generated in this study constitutes a new resource for genetic analysis in Atlantic herring significantly increasing the number of known transcripts as well as novel SNP and microsatellite markers.

Table 2. Results for SNP detection variables for predicting SNP assay conversion following a backward stepwise elimination procedure.

	B[a]	Wald[b]	df	P[c]
Asc_ind[d]	−0.165	4.67	1	**0.031**
ADS[e]	0.763	4.785	1	**0.029**
Constant	−0.378	1.464	1	0.226

[a]Regression coefficient for individual variable.
[b]Wald χ^2 statistic.
[c]associated probability.
[d]Number of ascertainment individuals with the minor allele.
[e]Assay Design Score. Significant p-values are shown in bold.
doi:10.1371/journal.pone.0042089.t002

Table 3. Results for SNP detection variables for predicting SNP assay polymorphism following a backward stepwise elimination procedure.

	B[a]	Wald[b]	df	P[c]
Asc_ind[d]	0.249	2.965	1	0.085
NSQ[e]	−0.111	7.321	1	**0.007**
Constant	0.935	21.137	1	0.000

[a]Regression coefficient for individual variable.
[b]Wald χ^2 statistic.
[c]associated probability.
[d]Number of ascertainment individuals with the minor allele.
[e]Neighbourhood Sequence Quality. Significant p-values are shown in bold.
doi:10.1371/journal.pone.0042089.t003

Sequence Assembly and SNP Detection

For next generation sequencing to be successfully applied to the development of genetic resources in non-model organisms, methodological issues must be addressed to optimise the procedures for each project. SNPs can be genome- or transcriptome derived and, in the latter case, selected from more abundant or rarer expressed transcripts; in addition, marker development is influenced by sequence depth and contig length due to the sequencing platform chosen and the complexity of the hypothesis to be investigated (i.e. smaller number of SNPs required for species identification analysis as compared to population genetic studies). The choice of sequencing platform should reflect the objective of a given study. While longer reads (e.g. 454 sequencing) are expected to improve contig assembly, more, but shorter, reads (e.g. Illumina sequencing) may be preferable in order to reduce detection of false positive SNPs from higher alignment depth, especially when an existing reference sequence is available. This study took advantage of the longer read lengths obtained with 454 sequencing in a *de novo* assembly of a reference scaffold for SNP discovery in herring. The clustering and assembly step is critical for SNP mining as it generates the reference for variant detection by mapping reads to the contig. Therefore, the absence of a reference genome or transcriptome poses a challenge for assessing the 'correctness' of a contig assembly, as potential mis-assemblies of sequence due to homologous or paralogous genes cannot be directly verified by back-mapping to the species-specific genome. Generally, cluster assembly with overly stringent parameters will lead to splitting sequences belonging together into more contigs, resulting in a higher number of shorter contigs with lower coverage depth. Whilst applying criteria that are overly relaxed will assemble reads from related genes or gene families into single contigs, resulting in a lower numbers of contigs that have a higher sequence depth, however this increases the likelihood of misidentifying polymorphisms between paralogous sequence variants (PSVs) as SNPs. Additionally, as no genome reference is available for Atlantic herring, the occurrence of PSVs cannot be assessed, this was probably the cause for the majority of ambiguous clustering that was subsequently seen in the SNPs.

For the SNP detection, the low sequence depth of the majority of contigs (Figure 3C) required relatively low criteria to be set (i.e. depth: four reads, redundancy: two observations of the minor

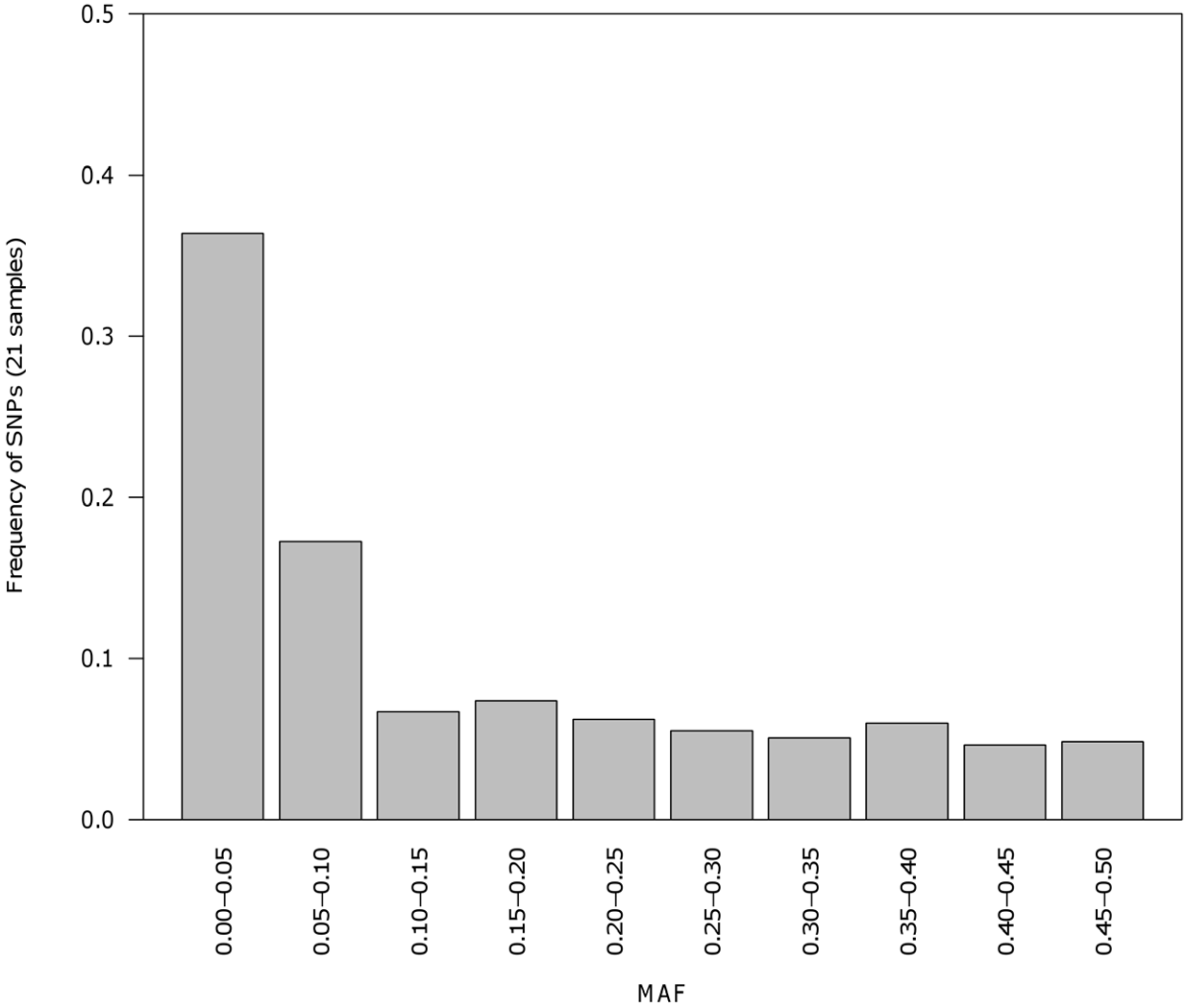


Figure 4. Minor allele frequency (MAF) distribution. The distribution of the MAF in 578 SNPs typed in 18 populations across the eastern herring distribution.
doi:10.1371/journal.pone.0042089.g004

allele). However, these low thresholds together with the sequencing of eight ascertainment individuals spanning the entire northeast Atlantic distribution of herring resulted in minimal ascertainment bias due to exclusion of low MAF SNPs (Figure 4). One expected result of the low depth and redundancy parameters is, however, the low conversion rate from the inflated number of candidate SNPs (identified due to sequencing errors). The 454 platform-specific challenge of resolving homopolymeric regions may further have compromised SNP detection by reducing assembly quality or calling false SNPs within these regions [41], but such an effect could not be assessed here due to the lack of a known reference sequence.

Table 4. Type and number of repeats of the microsatellites detected in the herring contigs using Msatcommander.

Type of repeat	Number of repeats					Total
	4–9	10–14	14–19	>19	Maximum	
Dinucleotide	4418	505	193	175	75	5291
Trinucleotide	829	35	9	2	36	875
Tetranucleotide	202	13	3	12	31	230
Pentanucleotide	43	1	1	0	17	45
Hexanucleotide	57	2	0	1	21	60
Total	5549	556	206	190	-	6501

doi:10.1371/journal.pone.0042089.t004

The use of transcriptome sequencing in this study has resulted in only a few per cent of the total genome being covered, but at a relatively high sequencing depth, thus limiting sequencing costs while achieving the number of SNPs required for custom-designed SNP assays. Additionally, transcriptome sequencing provides information about tissue-specific genes and their expression profile, which can be used to develop further tools for gene expression studies such as oligonucleotide microarray or RNA-seq approaches.

SNP Validation

The genotyping of 1,536 selected SNP assays performed with genomic DNA for a large panel of Atlantic herring samples from across the northeast Atlantic indicated that nearly 600 of the SNPs are polymorphic (37.6%). However, almost 49.3% of the candidate SNPs failed to work; due to either non-amplification

(18.9%), false positives (monomorphic loci) (13.1%) or ambiguous clustering (17.3%). Despite our attempt to screen for potential intron/exon splicing sites within flanking regions of all candidate SNPs using available reference genomes, only 41.9% of all queries matched equivalent sequences in at least one of the reference species. Thus, the presence of undetected introns may have constituted a major cause for genotyping failure [42]. Moreover, candidate SNPs that appeared monomorphic in the large-scale screening might either be the result of false-positive predictions or could indicate real, rare SNPs not present in the samples tested [7]. The purely *in silico* SNP detection method presented in this study may have a relatively low conversion rate to validated SNPs when compared to other methods. However, this method is still extremely competitive given a limited resource for marker development, once the time and cost associated with designing and ordering hundreds of primers, running validation PCRs, and additional Sanger sequencing for validation are considered (e.g. [39,40]). All of which would be in addition to the cost of genotyping the resulting 578 validated SNPs.

In order to reduce the number of erroneous SNP predictions, i.e. to increase the probability of an *in silico* detected SNP being a truly polymorphic site, further sequencing would lead to greater sequence depth of the contigs, allowing more stringent selection of SNP candidates. It has been shown for multiplexed re-sequencing that more than 90% of the variants can be detected correctly using next generation sequencing technologies when an average depth of at least 20 reads per base is achieved [43,44]. Increasing the average sequence depth will also be advantageous for identifying SNPs from rarely expressed genes. Another interesting approach, recently described by Ratan *et al.* [45], suggests a method to call SNPs without a reference genome sequence. SNP calling is performed whenever new sequences are added; thus, sequencing continues only as long as needed to identify an adequate number of candidate SNPs. The method is reported to work even when the sequence coverage is not sufficient for *de novo* assembly. Additionally, the use of next generation sequencing for analysing a restriction enzyme-generated DNA library (RRL and in particular RAD sequencing, for reviews see [46,47]) based on multiple tagged individuals now enables the fast discovery of thousands of SNPs in non-model organisms with no prior genome information [48,49]. However, one downstream problem identified with RAD-seq is that transferring the SNPs onto a high-throughput genotyping platform is difficult without a reference genome, as the majority of SNPs identified do not have the 60 bp flanking sequenced required for assay design. This has to some extent been solved using Paired End RAD (RAD-PE)[50], however the bioinformatic approaches for SNP discovery in RAD-PE contigs are still limited. Additionally, while RRL/RAD-seq approaches eliminate the problems encountered with intron/exon boundaries that are associated with transcriptome sequencing, these methods only consider random fragments of the entire genome, whereas our transcriptome based pipeline specifically targets expressed genes with an increased likelihood for detecting SNPs (e.g. non-synonymous substitutions) associated with genomic regions under selection. Such non-neutral SNPs are expected to provide high discriminatory power at the population level and will constitute a valuable forensic tool in future applications [47,51]. The combination of the coverage and SNP discovery rates obtained by RAD-seq, with the targeted reduction obtained by sequencing the transcriptome would potentially be a very powerful tool. However, it must be noted that due to the rapid rate of technical developments in the field, such as the increased read length and decreasing costs of existing platforms, and the potential of nano-sequencing technology, the best solution regarding platforms and methods to optimise the cost effectiveness for a specific application needs careful consideration.

When determining the predictive value of the SNP selection parameters for successful amplification of the *in silico* detected SNPs (*Conversion)*, as expected, a positive correlation was found with the Assay Design Score, i.e. the likelihood for designing successful primers around the SNP position. Unexpectedly, a negative correlation was found with number of ascertainment individuals for which the rare allele was observed, although the reasons behind this correlation are unclear. Overall, only very weak predictive variables for *Polymorphism* were identified, with only the neighbourhood sequence quality significantly explaining the negative correlation; as the number of mismatches in flanking regions increases, a predicted SNP is more likely to be a false positive. This increase in mismatches of an aligned region could be indicative of erroneous clustering, for example, PSVs or other sequences with differing genomic origin (this has for example also been seen for hake in a similar study [8]). The number of individuals with the minor allele in the ascertainment panel also showed a positive correlation with *Polymorphism.* While this parameter is less conclusive than for predicting *Conversion* rate, there is potentially a predictive role of this parameter for detecting true SNPs. Future SNP development efforts may reduce the false positive rate by applying relatively stringent thresholds for this variable (e.g. having at least 2 individuals with the minor allele represented in the SNP containing contig, although this will, of course, depend on the size of the ascertainment panel).

The two binomial logistic regression analyses were repeated with a reduced set of variables representing the strongest *a priori* candidates (the number of sequences aligned under the SNP position, the frequency of sequences with minor allele, the neighbourhood sequence quality, the Assay Design Score, and the outcome of the intron-exon boundary pipeline). This also allowed controlling for a potential bias from non-independent variables such as the two intron-exon and three minor allele related parameters. Results were largely congruent confirming Assay Design Score and neighbourhood sequence quality to be the most significant predictors of *Conversion* and *Polymorphism*, respectively.

The range of allele frequencies within the SNP panel suggests that the strategy of carefully selecting individuals to maximise the geographical, phenotypic and genetic diversity covered by the SNP development samples has been successful in minimising ascertainment bias.

Cross Species Amplification and Microsatellite Detection

A high proportion of detected SNPs also amplified single PCR products in Pacific herring albeit with a low polymorphism rate, which is as expected due to their development from conserved genomic regions. However, due to the small sample size (n = 4), this number is likely to be downwardly biased and a much higher proportion of SNPs may in fact be polymorphic and therefore prove useful in this species. As expected from the phylogenetics of these species, the proportions of SNP amplification and polymorphism were lower in the anchovy. Additionally, our sequencing effort has led to the discovery of a large resource of microsatellite markers, 36% of which have primers successfully designed (Table S3). These include both neutral loci and loci that are physically linked to SNPs representing genomic regions that have been shown to be under directional selection [38]. Another attribute of multi-allelic microsatellite markers when studying adaptive genetic variation is the increased statistical power for detecting balancing selection compared to bi-allelic markers (such as SNPs, e.g. [52]), and also for applications such as parental assignment.

Conclusion

Our approach of applying barcoding and multiplexing individuals for large-scale *in silico* mining of transcriptome sequences seems to be a very appropriate strategy to develop new SNP markers in non-model species as it does not require costly and time-intensive re-sequencing of target amplicons necessitating prior knowledge and availability of genome sequence information. However, the purely *in silico* based SNP detection comes with a trade off in the form of an expectedly lower conversion rate in the final genotyping assay [53]. The resultant resources will be of value in on-going analyses of population structuring and stock dynamics, assays of adaptive variation, and for enhancing the scope of microsatellite-based studies.

Supporting Information

Figure S1 Analysis pipeline. The path on the left of the figure illustrates the pipeline for the genomic approach, where herring transcripts are directly compared with five reference genomes. The path on the right of the figure shows the pipeline for the transcriptomic approach, where herring transcripts are first compared to the transcriptome of the five reference species. Hits were then subsequently matched to the corresponding genomes of the same species (see text for more details).
(TIF)

Table S1 Number of contigs and singletons annotated using a range of fish and human reference resources and databases.
(XLSX)

Table S2 List of the 578 validated polymorphic SNPs found in this study, including the 120 bp flanking region, with the two SNP alleles in brackets. Also global estimates of observed (Ho) and expected heterozygosity (He) in the four ascertainment populations for each SNP. The S/NS column denotes whether a SNP was either synonymous (S) or non-synonymous (NS) with NA designating SNPs with no contig match in the BLAST search (see text for more details).
(XLSX)

Table S3 List of the microsatellites for which primers were successfully designed, along with up to 200 bases flanking sequence.
(XLSX)

Acknowledgments

We would like to thank all the members of the FishPopTrace Consortium for their input.

Sampling was made possible by the generous collaboration of Eero Aro, Philip Coupland, Geir Dahle, Audrey Geffen, Thomas Gröhsler, Birgitta Krischansson, Ciaran O'Donnell, Henn Ojaver, Guðmundur Óskarsson, Iain Penny, Jukka Pönni, Fausto Tinti, Veronique Verrez-Bagnis Phil Watts, and Miroslaw Wyszynski. We thank Pernille K. Andersen (Aarhus University, Denmark) for sequencing and library management.

Author Contributions

Conceived and designed the experiments: GRC FP SJH DB MIT LB RO GEM JvH FPT Consortium. Performed the experiments: FP RON RO SJH MTL. Contributed reagents/materials/analysis tools: FPT Consortium. Wrote the paper: SJH MTL MIT DB GRC FP. Carried out in silico analyses: FP RON SJH MTL MIT MB JvH GEM AC. Analyzed genotype data: SJH MTL DB MIT. Carried out statistical analysis: SJH MTL DB LB MB.

References

1. Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, et al. (2011) Application of SNPs for population genetics of non-model organisms: new opportunities and challenges. Molecular Ecology Resources 11(S1): 123–136.
2. Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, et al. (2010) Adaptation genomics: the next generation. Trends in Ecology & Evolution 25: 705–712.
3. Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009) Population genomics of marine fishes: Identifying adaptive variation in space and time. Molecular Ecology 18: 3128–50.
4. Waples RS, Punt AE, Cope JM (2008) Integrating genetic data into management of marine resources: how can we do it better? Fish and Fisheries 9: 423–449.
5. Corrigendum to Council Regulation (2008) (EC) No 1005/2008 of 29 September 2008 establishing a Community system to prevent, deter and eliminate illegal, unreported and unregulated fishing, amending Regulations (EEC) No 2847/93, (EC) No 1936/2001 and (EC) No 601/2004 and repealing Regulations (EC) No 1093/94 and (EC) No 1447/1999. Official Journal of the European Union, L 286.
6. FAO Fisheries and Aquaculture Report No. 973 (FIPI/R973) Rome, 31 January–4 February 2011. Report of the twenty-ninth session of the committee on Fisheries. Available: http://www.fao.org/docrep/014/i2281e/i2281e00.pdf. Accessed 2011 Feb 7.
7. Hubert S, Higgins B, Borza T, Bowman S (2010) Development of a SNP resource and a genetic linkage map for Atlantic cod (Gadus morhua). BMC Genomics 11: 191.
8. Milano I, Babbucci M, Panitz F, Ogden R, Nielsen RO, et al. (2011) Novel tools for conservation genomics: Comparing two high-throughput approaches for SNP discovery in the transcriptome of the European hake. PLoS ONE 6: e28008.
9. Li R, Fan W, Tian G, Zhu H, He L, et al. (2010) The sequence and de novo assembly of the giant panda genome. Nature 463: 311–317.
10. Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, et al. (2011) The genome of *Theobroma cacao*. Nature Genetics 43: 101–108.
11. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrom M, et al. (2011) The genome sequence of Atlantic cod reveals a unique immune system. Nature 477: 207–210.
12. Sánchez CC, Smith TP, Wiedmann RT, Vallejo RL, Salem M, et al. (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. BMC Genomics: 10: 559.
13. Kerstens HH, Crooijmans RP, Veenendaal A, Dibbits BW, Chin-A-Woeng TF, et al. (2009) Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. BMC Genomics: 10: 479.
14. van Bers NE, van Oers K, Kerstens HH, Dibbits BW, Crooijmans RP, et al. (2010) Genome-wide SNP detection in the great tit *Parus major* using high throughput sequencing. Molecular Ecology: 19(S1): 89–99.
15. McPherson AA, Stephenson RL, O'Reilly PT, Jones MW, Taggart CT (2001) Genetic diversity of coastal Northwest Atlantic herring populations: implications for management. Journal of Fish Biology 59: 356–370.
16. Bekkevold D, Andre C, Dahlgren TG, Clausen LAW, Torstensen E, et al. (2005). Environmental correlates of population differentiation in Atlantic herring. Evolution 59: 2656–2668.
17. Jørgensen HBH, Hansen MM, Bekkevold D, Ruzzante DE, Loeschcke V (2005) Marine landscapes and population genetic structure of herring (*Clupea harengus* L.) in the Baltic Sea. Molecular Ecology 14: 3219–3234.
18. Larsson LC, Laikre L, Palm S, Andre C, Carvalho GR, et al. (2007) Concordance of allozyme and microsatellite differentiation in a marine fish, but evidence of selection at a microsatellite locus. Molecular Ecology 16: 1135–1147.
19. Gaggiotti OE, Bekkevold D, Jorgensen HBH, Foll M, Carvalho GR, et al. (2009) Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. Evolution 63: 2939–2951.
20. Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, et al. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS ONE 2: e197.
21. Lavoué S, Miya M, Saitoh K, Ishiguro NB, Nishida M (2007) Phylogenetic relationships among anchovies, sardines, herrings and their relatives (Clupeiformes), inferred from whole mitogenome sequences. Molecular Phylogenetics and Evolution 43: 1096–1105.
22. Smit AFA, Hubley R, Green P (1996) *RepeatMasker Open-3.0*. Available http://www.repeatmasker.org. version open-3.2.7 with RM database version 20090120.
23. Huang XQ, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Research 9: 868–877.
24. Marth GT, Korf I, Yandell MD, Yeh RT, Gu ZJ, et al. (1999) A general approach to single-nucleotide polymorphism discovery. Nature Genetics 23: 452–456.
25. Bai J, Li Q, Cong RH, Sun WJ, Liu J, et al. (2011) Development and characterization of 68 EST-SSR markers in the Pacific oyster, *Crassostrea gigas*. Journal of the World Aquaculture Society 42(3): 444–455.

26. Pashley CH, Ellis JR, McCauley DE, Burke JM (2006). EST databases as a source for molecular markers: lessons from *Helianthus*. Journal of Heredity 97: 381–388.
27. Faircloth BC (2008) MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. Molecular Ecology Resources 8: 92–94.
28. Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Bioinformatics Methods and Protocols: Methods in Molecular Biology (eds Krawetz S, Misener S). Humana Press, Totowa, NJ.
29. Castoe TA, Poole AW, Gu W, Jason de Koning AP, Daza JM, et al. (2010) Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. Molecular Ecology Resources10: 341–347.
30. Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, et al. (2003) Highly parallel SNP genotyping. Cold Spring Harbor Symposia on Quantitative Biology 68: 69–78.
31. Li C, Ortí G (2007) Molecular phylogeny of Clupeiformes (Actinopterygii) inferred from nuclear and mitochondrial DNA sequences. Molecular Phylogenetics and Evolution 44: 386–398.
32. Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes 6: 288–295.
33. Rousset F (2008) GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. Molecular Ecology Resources 8(1): 103–106.
34. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29: 1165–1188.
35. Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Molecular Biology and Evolution 27: 2534–2547.
36. Rosenblum EB, Novembre J (2007) Ascertainment bias in spatially structured populations: A case study in the eastern fence lizard. Journal of Heredity 98: 331–336.
37. Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166: 351–372.
38. Limborg MT, Helyar SJ, de Bruyn M, Taylor MI, Nielsen EE, et al. (2012) Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). Molecular Ecology doi: 10.1111/j.1365–294X.2012.05639.x.
39. Geraldes A, Pang J, Thiessen N, Cezard T, Moore R, et al. (2011) SNP discovery in black cottonwood (*Populous trichocarpa*) by population transcriptome resequencing. Molecular Ecology Resources 11(S1): 81–92.
40. Seeb JE, Pascal CE, Grau ED, Seeb LW, Templin WD, et al. (2011) Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids. Molecular Ecology Resources 11(S1): 335–348.
41. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
42. Wang SL, Sha ZX, Sonstegard TS, Liu H, Xu P, et al. (2008) Quality assessment parameters for EST-derived SNPs from catfish. BMC Genomics 9: 450.
43. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, et al. (2008) Identification of genetic variants using bar-coded multiplexed sequencing. Nature Methods 5: 887–893.
44. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biology 10: R32.
45. Ratan A, Yu Z, Hayes VM, Schuster SC, Miller W (2010) Calling SNPs without a reference sequence. BMC Bioinformatics 11: 130.
46. Ogden R (2011) Unlocking the potential for genomic technologies for wildlife forensics. Molecular Ecology Resources 11(S1): 109–116.
47. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Review Genetics 12: 499–510.
48. Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Molecular Ecology Resources 11(S1): 117–122.
49. Van Tassell CP, Smith TL, Matukumalli LK, Taylor JF, Schnabel RD, et al. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods 5: 247–252.
50. Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011) Local *De Novo* Assembly of RAD Paired-End Contigs Using Short Sequencing Reads. PLoS ONE 6(4): e18561.
51. Nielsen E, Cariani A, Mac Aoidh E, Maes G, Milano I, et al. (2012) Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. Nature Communications 3: 851 doi: 10.1038/ncomms1845.
52. Narum SR, Hess JE (2011) Comparison of F-ST outlier tests for SNP loci under selection. Molecular Ecology Resources 11(S1): 184–194.
53. Lepoittevin C, Frigerio JM, Garnier-Géré P, Salin F, Cervera MT, et al. (2010) *In vitro* vs *in silico* detected SNPs for the development of a genotyping array: what can we learn from a non-model species? PLoS One 5: e11034.

Chapter 7

Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*)

Published in *Molecular Ecology*

MOLECULAR ECOLOGY

Molecular Ecology (2012) doi: 10.1111/j.1365-294X.2012.05639.x

Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*)

MORTEN T. LIMBORG,*[1] SARAH J. HELYAR,†‡[1] MARK DE BRUYN,† MARTIN I. TAYLOR,† EINAR E. NIELSEN,* ROB OGDEN,§ GARY R. CARVALHO†, FPT CONSORTIUM and DORTE BEKKEVOLD*

*National Institute of Aquatic Resources, Technical University of Denmark, Vejlsøvej 39, DK-8600 Silkeborg, Denmark, †Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Environment Centre Wales, Bangor University, Bangor LL57 2UW, UK, ‡Matís, Vínlandsleið 12, 113 Reykjavík, Iceland, §TRACE Wildlife Forensics Network, Royal Zoological Society of Scotland, Corstorphine Road, Edinburgh EH12 6TS, UK

Abstract

High gene flow is considered the norm for most marine organisms and is expected to limit their ability to adapt to local environments. Few studies have directly compared the patterns of differentiation at neutral and selected gene loci in marine organisms. We analysed a transcriptome-derived panel of 281 SNPs in Atlantic herring (*Clupea harengus*), a highly migratory small pelagic fish, for elucidating neutral and selected genetic variation among populations and to identify candidate genes for environmental adaptation. We analysed 607 individuals from 18 spawning locations in the northeast Atlantic, including two temperature clines (5–12 °C) and two salinity clines (5–35‰). By combining genome scan and landscape genetic analyses, four genetically distinct groups of herring were identified: Baltic Sea, Baltic–North Sea transition area, North Sea/British Isles and North Atlantic; notably, samples exhibited divergent clustering patterns for neutral and selected loci. We found statistically strong evidence for divergent selection at 16 outlier loci on a global scale, and significant correlations with temperature and salinity at nine loci. On regional scales, we identified two outlier loci with parallel patterns across temperature clines and five loci associated with temperature in the North Sea/North Atlantic. Likewise, we found seven replicated outliers, of which five were significantly associated with low salinity across both salinity clines. Our results reveal a complex pattern of varying spatial genetic variation among outlier loci, likely reflecting adaptations to local environments. In addition to disclosing the fine scale of local adaptation in a highly vagile species, our data emphasize the need to preserve functionally important biodiversity.

Keywords: genome scan, haemoglobin, heat shock protein, local adaptation, salinity, single nucleotide polymorphism

Received 21 December 2011; revision received 30 March 2012; accepted 14 April 2012

Introduction

Local adaptation can evolve only if the strength of divergent selection overrides random genetic drift and the homogenizing effect of gene flow among populations (Kawecki & Ebert 2004). These premises suggest that the occurrence of local adaptation should be rare in high gene flow species such as many marine organisms (Palumbi 1994; Conover *et al.* 2006). In contrast, large effective population sizes (N_e) should enhance response to selection, and local selective pressures may be substantial considering the often immense environmental heterogeneity experienced by widely distributed marine species. A recent simulation-based study showed that

Correspondence: Morten T. Limborg, Fax: +45 35 88 31 50; E-mail: mol@aqua.dtu.dk

[1]Joint first authors.

even in the face of considerable gene flow, environmental heterogeneity may cause disruptive selection and result in local adaptation (Yeaman & Whitlock 2011). Accordingly, expectations are that genes and linked regions under the influence of divergent selection will show elevated differentiation, in comparison with selectively 'neutral' gene regions. Until now, genomic studies of high gene flow marine fish have been mostly restricted to Atlantic cod (*Gadus morhua*) (Moen *et al.* 2008; Nielsen *et al.* 2009b; Bradbury *et al.* 2010), while for other fishes, inference of genic selection has often been made from a single or few candidate genes (Hemmer-Hansen *et al.* 2007a; Gaggiotti *et al.* 2009; Larmuseau *et al.* 2009).

The task of identifying signatures of natural selection in nonmodel species has been constrained by often limited numbers of (usually) neutral genetic markers (Hauser & Seeb 2008). Next-generation sequencing (NGS) technologies have facilitated the development of large transcriptome-derived marker panels, effectively increasing the chance of detecting natural selection by studying functional genetic variation, which is expected to be more directly affected by natural selection (Allendorf *et al.* 2010). The increased genomic coverage further improves the chance of detecting loci affected by divergent selection from neutrally evolving sites by applying genome scan approaches (Beaumont 2005; Storz 2005). The adaptive significance of actual outlier loci is often elusive because they may not be the direct target of selection but rather exhibit hitchhiking with genes under selection (Maynard Smith & Haigh 1974). However, the combination of insights from known gene functions, landscape effects (Manel *et al.* 2003), replicated patterns across independent environmental clines (Schmidt *et al.* 2008) and previous findings provides stronger evidence for adaptive roles of outlier loci (Vasemägi & Primmer 2005).

Despite a predominant picture of weak population structure in most marine fishes (Ward *et al.* 1994), genomic regions under divergent selection may be more prevalent than hitherto anticipated (Nielsen *et al.* 2009a). In the present study, we use the Atlantic herring (*Clupea harengus*; hereafter 'herring') as a model to investigate spatially explicit genomic variation in a marine organism characterized by high gene flow and large effective population size (N_e). Herring is a small, highly migratory pelagic fish distributed throughout heterogeneous environments in large parts of the North Atlantic. Local populations exhibit large differences in demographic and life history parameters including growth, spawning season and migratory behaviour (Iles & Sinclair 1982; Aro 1989). Outside spawning seasons, several populations undergo long-distance migrations to communal feeding areas (e.g. Ruzzante *et al.* 2006), suggesting ample opportunities for dispersal and gene flow. In some areas, a combination of high gene flow and large N_e among herring populations presumably impedes genetic detection of local demes using neutral markers (Mariani *et al.* 2005). However, for other geographical regions, significant genetic structuring is evident, especially across the strong environmental cline separating the fully marine North Sea from the brackish Baltic Sea (Bekkevold *et al.* 2005) as well as weak, but statistically significant, patterns within the Baltic Sea (Jørgensen *et al.* 2005). More recently, signatures of selection have also been demonstrated in herring (Larsson *et al.* 2007; Gaggiotti *et al.* 2009; Andre *et al.* 2011), but these studies focus on comparisons between North Sea and Baltic Sea herring for a single microsatellite locus. Thus, despite many population genetic studies on herring, the geographical scale and pattern of adaptive divergence at genomic levels remains largely unknown.

We investigated the spatial and genomic scales at which herring populations are likely to exhibit adaptation to local environments. We conducted comprehensive sampling of herring spawning populations throughout the northeastern Atlantic, and across several environmental gradients, and applied a statistical genome scan approach to transcriptome-derived single-nucleotide polymorphism (SNP) markers. To assess the robustness of loci under selection, we use two different 'outlier tests' for identifying gene regions exhibiting statistical evidence of predominantly either neutral or divergent selection processes. Furthermore, we use a complementary 'landscape genetics' approach to identify loci under divergent selection in relation to key environmental parameters. Findings are discussed in relation to the prospects and significance of detecting functional biodiversity in high gene flow taxa through exploring genes subject to local adaptive evolution in the oceans.

Materials and methods

Samples

Twenty-one samples were collected from scientific surveys and commercial fishing vessels, representing 18 locations spanning the majority of the species' east Atlantic distribution (Fig. 1). Three samples represented temporal (range = 6–10 years) replicates within locations. Populations were targeted during the spawning season at known spawning grounds and mainly comprised spawning (ripe and running) individuals. Samples spanned latitudinal clines (reflecting temperature) both in the North Sea/North Atlantic and in the Baltic Sea (Fig. 1). Samples also covered longitudinal clines (corresponding with two low-salinity environments): one

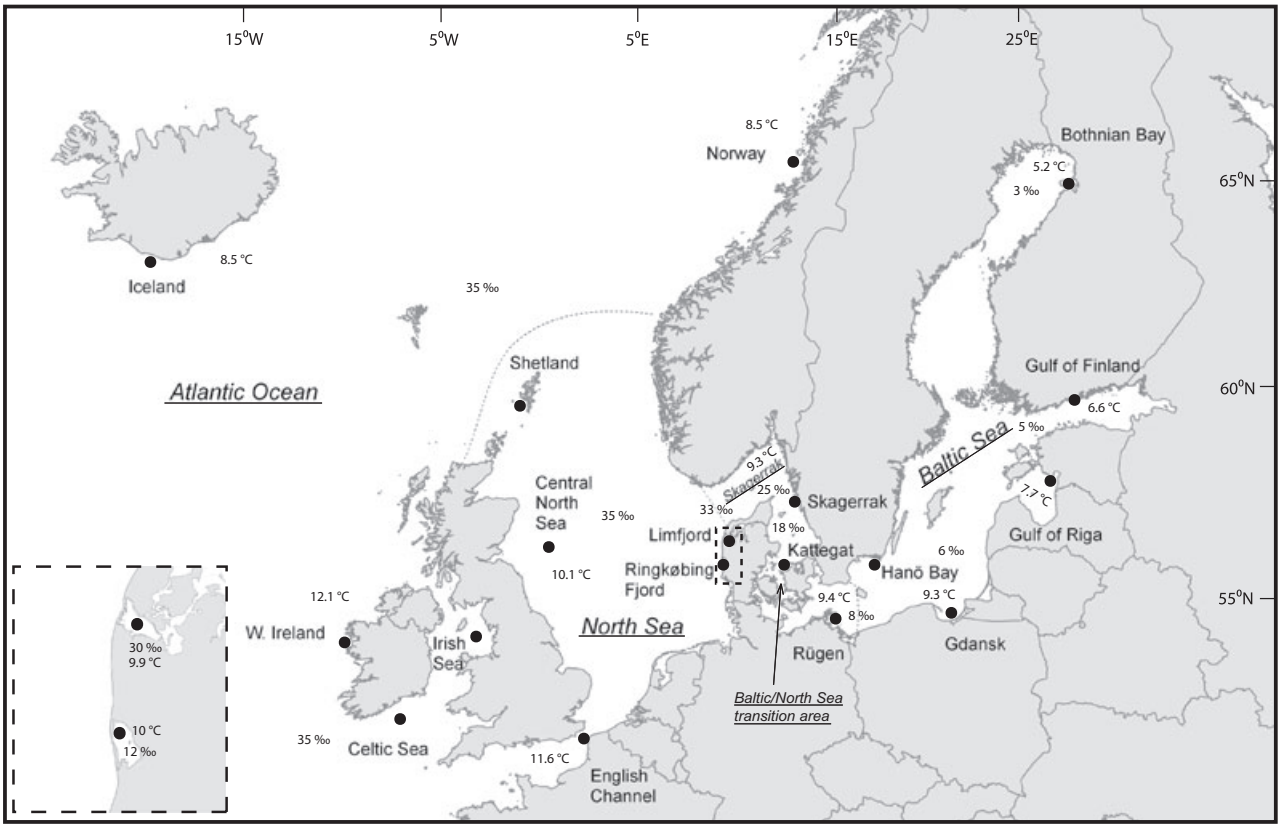


Fig. 1 Distribution of sampled locations. Average annual surface temperature (°C) and salinities (‰) are given throughout the distribution, and major regional areas are denoted in italics.

spanning the North Sea/British Isles into the Baltic Sea which is an isolated brackish sea only receiving saline waters from the North Sea through the narrow Danish Straits and one going from the North Sea/British Isles with high-salinity coastal locations, changing to more brackish spawning locations in the Ringkøbing Fjord draining into the eastern North Sea and separated from the low-saline Baltic Sea cline (Fig. 1). Spawning times of herring differ among local populations (Cushing 1967) and collections reflect this, as spring-, autumn- and winter-spawning populations are all represented (Table 1).

Molecular analyses and genotyping

DNA was extracted from gill, muscle or fin tissue stored in 96% ethanol using the E.Z.N.A. Tissue DNA kit (Omega Bio-Tek, Norcross, GA, USA) following the manufacturer's protocol. A NanoDrop Spectrophotometer (Thermo Fisher Scientific Inc.) was used to ensure adequate quality and quantity of DNA prior to genotyping. A total of 762 individuals were screened for a panel of 310 SNPs (Table S1, Supporting information). These SNPs were selected from a larger collection of 578 SNPs (Helyar *et al.* 2012) based on robust clustering of genotypes (see below) and minimization of the number of loci affected by very strong linkage disequilibrium. Genotyping was performed using a custom Illumina Golden Gate® Assay (Fan *et al.* 2003) in Sentrix Array Matrix (SAM) format on the iScan platform. SNPs were developed from an ascertainment panel of eight individuals representing all major geographical regions studied here (i.e. Baltic Sea, English Channel, North Sea and North Atlantic) (Helyar *et al.* 2012), thus minimizing ascertainment bias (Rosenblum & Novembre 2007). Overall, 29 SNPs were discarded due to ambiguous clustering (Table S1, Supporting information), and of the remaining 281 SNPs, 70% were annotated using BLASTN (NCBI) (Helyar *et al.* 2012). Genotyping data were visualized and analysed using the GENOMESTUDIO Data Analysis Software package (1.0.2.20706; Illumina Inc.). One specimen was independently re-genotyped 12 times, which allowed the estimation of an overall genotyping error of 1.57% across all loci and samples. Specimens with low call rates (<90% loci genotyped) were discarded, leading to 607 individuals genotyped for a total of 281 SNPs (Table S1, Supporting information) in 21 sample collections (n = 17–39).

Table 1 Sample information including peak spawning times for the populations sampled

Geographical region	Sample location	Year	Month	Peak spawning	*n*	Latitude	Longitude	Genetic cluster*	H_e	H_o	Environmental conditions†			
											SST	SST. spawn	PSU	PSU.spawn
North Atlantic	Norway[1,2]	2009	Sep	Mar	31	65.54‡	11.26‡	North Atlantic	0.30	0.30	8.5	6.0	33.7	33.9
	Iceland[1,2]	2009	Jul	May–Sep	34	63.62	−19.62	North Atlantic	0.30	0.30	8.5	10.9	34.9	34.5
North Sea/British Isles	Shetland[1,2]	2009	Aug	Aug	34	60.35	−2.72	North Sea/British Isles	0.30	0.30	10.5	11.7	35.2	35.2
	W. Ireland[1,4,5]	2004	Nov	Nov	28	53.90	−10.36	North Sea/British Isles	0.31	0.32	12.1	13.0	35.2	35.2
	Celtic Sea[1]	2008	Oct	Oct	39	51.24	−8.26	North Sea/British Isles	0.30	0.30	12.4	12.8	34.9	35.0
	Irish Sea[1,4,5]	2009	Sep	Sep	36	54.03	−4.07	North Sea/British Isles	0.30	0.31	10.8	13.3	33.5	33.6
	English Channel[1]	1999	Nov	Nov–Jan	17	50.81	1.57	North Sea/British Isles	0.30	0.31	11.6	8.5	34.9	34.9
	English Channel[1,2]	2009	Jan	Nov–Jan	36	50.81	1.57	North Sea/British Isles	0.30	0.31	11.6	8.5	34.9	34.9
	Central North Sea[1,2,4,5]	2009	Aug	Aug	30	56.43	0.20	North Sea/British Isles	0.30	0.30	10.1	11.4	34.9	34.9
	Ringkøbing Fjord[1,5]	2009	Apr	Apr	33	55.97	8.24	B/NS transition area	0.30	0.31	10.0	12.5	9.1	9.0
	Limfjord[1]	2009	Apr	Apr	33	56.60	8.35	B/NS transition area	0.30	0.31	9.9	11.5	30.8	30.7
B/NS transition area	Skagerrak[1,4]	2009	Mar	Apr	36	57.40	11.40	North Sea/British Isles	0.30	0.31	9.3	5.6	24.9	23.8
	Kattegat[1,4]	2003	Apr	Apr	23	55.73	11.37	B/NS transition area	0.29	0.31	9.4	6.0	18.7	17.1
	Rügen[1]	2003	Apr	Mar–Apr	19	54.21	13.62	B/NS transition area	0.30	0.31	9.4	5.3	8.0	8.1
	Rügen[1,4]	2009	Mar	Mar–Apr	36	54.21	13.62	B/NS transition area	0.30	0.30	9.4	5.3	8.0	8.1
Baltic Sea	Hanö Bay[1,3]	2002	Apr	Apr	24	55.57	15.18	Baltic Sea	0.31	0.32	8.6	7.4	7.5	7.5
	Gdansk[1,3,4]	2009	Mar	Mar	17	54.37	19.67	Baltic Sea	0.27	0.30	9.3	4.5	7.3	7.2
	Gulf of Riga[1]	2002	May	May–Jun	17	57.83	22.83	Baltic Sea	0.29	0.32	7.7	6.4	5.5	5.5
	Gulf of Riga[1,3,4]	2008	Jun	May–Jun	27	57.83	22.83	Baltic Sea	0.27	0.29	7.7	6.4	5.5	5.5
	Gulf of Finland[1,3]	2009	May	May	24	60.40	26.70	Baltic Sea	0.27	0.28	6.6	4.8	5.5	5.3
	Bothnian Bay[1,3]	2009	Jun	Jun	33	65.05	24.58	Baltic Sea	0.30	0.30	5.2	8.1	2.9	2.6

Numbers after sample location names denote samples included in (1) Global and regional analyses, (2) North Sea/North Atlantic latitudinal cline, (3) Baltic Sea latitudinal cline, (4) North Sea/British Isles—Baltic Sea longitudinal cline and (5) North Sea/British Isles—Ringkøbing Fjord longitudinal cline. Also shown for each sample is the most likely of four clusters as inferred from STRUCTURE analysis (B/NS = Baltic Sea/North Sea). Expected (H_e) and observed (H_o) heterozygosities are shown for each population. Environmental conditions used in landscape genetic analyses are shown for each sample location as SST (annual mean sea surface temperature), SST.spawn (spawning period mean sea surface temperature), PSU (sea surface salinity) and PSU.spawn (spawning period mean sea surface salinity) (See text for more details).

*Based on STRUCTURE analysis including the 'full' marker set and for $K = 4$ (see text). For each sample, membership to the genetic cluster receiving highest support is shown, disregarding that some samples showed high levels of admixture (e.g. Skagerrak, Fig. 3).

†Temperature and salinity data sources for Ringkøbing Fjord and Limfjord (http://www.dmu.dk/vand/havmiljoe/mads/ctd/data) and all other samples (http://www.ices.dk/ocean/data/surface/surface.htm).

‡This population was sampled off the major spawning ground at: 70.06 N and 16.90 E; however, preliminary otolith analyses showed that this sample represents the major Norwegian spring-spawning herring population (FishPopTrace consortium), and we use coordinates for the major spawning area in the landscape genetics analyses.

Summary statistics

Within each population, loci were tested for departure from Hardy–Weinberg proportions (HWE) using ARLEQUIN 3.5 (Excoffier & Lischer 2010) with a Markov Chain (MC) of length 10^6 and 100 000 dememorizations. A false discovery rate (FDR) was calculated to correct for multiple testing using the approach by Benjamini & Yekutieli (2001). Linkage disequilibrium was tested for each marker pair in all samples with GENEPOP 4.0 (Raymond & Rousset 1995) (10 000 dememorizations, 100 batches and 5000 iterations), and the results were corrected for multiple testing as above. For each population, estimates of expected (H_e) and observed (H_o) heterozygosities were obtained using GENALEX 6.4 (Peakall & Smouse 2006).

Outlier analyses

Two independent methods were used to identify putative loci under selection. ARLEQUIN v3.5 (Excoffier & Lischer 2010) utilizes coalescent simulations to generate a null distribution of *F*-statistics, with *P*-values conditioned on observed levels of heterozygosities across loci (Excoffier *et al.* 2009). Excoffier *et al.* (2009) demonstrated that the hierarchical island model produces fewer false positives than the finite island model for species exhibiting spatial population structure. For comparison, we tested both models. The hierarchical island model was implemented by grouping population samples according to the genetic clustering analyses (Table 1), as follows: (i) the Baltic Sea, (ii) the Baltic/North Sea transition area, (iii) the North Sea/British Isles and (iv) the North Atlantic. For all analyses, the settings were 10 000 simulations, 100 demes per group, and 10 groups. Loci that fell outside the 95% quantile were regarded as candidates for selection. BAYESCAN v2.01 (Foll & Gaggiotti 2008) measures the discord between global and population-specific allele frequencies (based on F_{ST} coefficients). While this method does not take into account the population structure, simulations have shown BAYESCAN to have lower type I and II errors than ARLEQUIN (Narum & Hess 2011). Log10 values of the posterior odds (PO) >0.5 and 2.0 were taken as 'substantial' and 'decisive' evidence for selection (Jeffreys 1961). An advantage of the posterior probability approach is that it directly allows for control of the FDR; here, the FDR was set at 0.05 and 0.01, adjusting the log10(PO) significance thresholds corresponding to the 0.5 and 2.0 values considered before correction. To compare global and regionally based signatures of selection, we performed global (21 population samples) genome scans using both software packages as detailed above. Based on the combined inference from these global genome scans, each SNP was categorized as either an 'outlier' (if it came out as such with either one or both of the programs) or 'neutral' (if it showed no indication of outlier behaviour with either program). We then constructed two data sets: one including both outlier and neutrally behaving SNP loci (referred to as the 'full' marker set) and one where all loci detected as outliers in global tests were removed (referred to as 'neutral' marker set). To further increase support for potential outliers in relation to environmental adaptation, we performed local genome scans to focus on the two separate regional temperature clines and two salinity clines (see Table 1 for samples included).

Population structure

For the 'neutral' marker set, temporal stability between replicates for three locations (Table 1) was assessed through pairwise F_{ST} analyses (following Weir & Cockerham 1984) using the *Fstat* function implemented in GENELAND (Guillot *et al.* 2005) conducted in the program *R* (http://cran.r-project.org). Temporal samples not exhibiting significant differentiation ($\alpha = 0.05$) were pooled within locations for subsequent analyses. F_{ST} was computed between all pairs of samples using both 'neutral' and 'full' marker sets. For all comparisons, significance was tested by permuting individuals 10 000 times among samples followed by correction for multiple tests using the FDR ($\alpha = 0.05$) according to Benjamini & Yekutieli (2001). The statistical power of the 'neutral' marker set for detecting genetic differentiation was assessed using POWSIM (Ryman & Palm 2006). By defining a given effective population size (N_e), POWSIM simulates genetic drift within two independent populations for t generations. N_e was set to 10 000 (the maximum allowed) and t varied among simulations to obtain a range of known F_{ST} values (0.00–0.02) between two hypothetical populations. Hereafter, 40 individuals were sampled from each population, and the null hypothesis of genetic homogeneity between samples was tested using a chi-square test. Repeating this procedure 1000 times allowed the assessment of the statistical power as the proportion of significant outcomes for each level of F_{ST}.

To infer the number of major genetic clusters, we used the Bayesian MCMC clustering approach implemented in STRUCTURE 2.3.1 (Pritchard *et al.* 2000). This model clusters all individuals into a predefined number of clusters (K) by minimizing overall deviation from HW and linkage equilibrium within clusters. Considering previous findings of high levels of gene flow in herring, we used the admixture model with correlated allele frequencies to reflect the most likely pattern of population connectivity. Also, due to the sampling

design, we allowed the model to include prior information on sampling location (Hubisz *et al.* 2009). Ten independent trials were run for each predefined K value, with $K = 1$–10. We used a burn-in of 10 000 iterations followed by 100 000 MCMC repetitions, and consistency of the three most likely K estimates was confirmed by longer chains of 100 000 burn-in and 500 000 final repetitions. In order to identify the most likely number of genetic clusters, also considering a sound biological interpretation, we initially considered both raw probability values of $\ln P(X \mid K)$ given by the program, and the ΔK estimate (Evanno *et al.* 2005). Where two models with consecutive K values could not be statistically distinguished, we performed hierarchical AMOVA using the locus-by-locus approach and 10 000 permutations in ARLEQUIN 3.5 (Excoffier & Lischer 2010), using both the 'neutral' and 'full' marker sets.

In addition, barplots of individual admixture proportions were visually inspected to infer the biologically most meaningful value of K. For example, if increasing K by one simply added a new cluster equally represented by all individuals in the data, as opposed to the break-up of existing clusters forming a new more or less admixed cluster, the lower value of K would be considered more biologically realistic.

Environmental associations with genetic variation

To test for association between specific gene regions and environmental or landscape parameters, we applied the Bayesian approach implemented in BAYENV (Coop *et al.* 2010). This approach takes into account the effect of underlying (neutral) population structure by first estimating a covariance matrix based on neutral markers, which is subsequently used to control for demographic variation when testing landscape- and locus-specific correlations in a Bayesian framework (see Coop *et al.* 2010). For this, we estimated a neutral covariance matrix based on the 'neutral' marker set. Results are given as a Bayes factor (BF) for each landscape variable and SNP locus correlation. This BF represents a ratio of the posterior likelihoods of a model where the landscape parameter has a significant effect on the locus, over an alternative model with no effect of the tested variable. We considered log10(BF) values above 1.5 as 'very strong' evidence (Jeffreys 1961) for an effect of the tested landscape/environmental factor (or any correlated factors) on the observed SNP allele distribution. The following landscape/environmental parameters were considered: (i) latitude, (ii) longitude, (iii) mean annual surface salinity, (iv) mean annual surface temperature, (v) mean spawning period surface salinity and (vi) mean spawning period surface temperature (Table 1). The latter two parameters were tested based on the assumption that mortality selection is expected to be most important during the egg (7–14 days) and larval (*c.* 2 months) phases, when natural mortality is highest (Dahlberg 1979). Estimates for temperature and salinity were calculated as the mean value over periods ranging from 20 to 120 years (depending on data availability) for all months (annual means) or for the 3 months following the midpoint of the spawning period (data and sources are listed in Table 1). As all pairwise genetic comparisons between temporally replicated samples were nonsignificant (see Results), all within-population genotypes across years were pooled for these analyses. To test for relationships between selected genetic variation and environment across both global and local scales, we performed a global analysis including all 18 samples as well as four regionally based analyses (regions defined as per regional genome scans; Table 1). To further rule out potential false positive correlations resulting from covarying isolation-by-distance (IBD) effects, we performed partial Mantel tests (Legendre & Legendre 1998) of locus-specific pairwise F_{ST} matrices and environmental distances for all loci correlating with temperature and salinity while controlling for geographical distance (shortest waterway) using the NCF (spatial nonparametric covariance functions) package in R (http://cran.r-project.org/web/packages/ncf/index.html) and running 1000 simulations to test for significance.

Results

Summary statistics

After removing loci that were monomorphic within a population, a total of 5541 tests for deviation from HWE were performed across all samples. Prior to and following correction for multiple testing (FDR = 5%), 159 (2.6%) and 19 (0.3%) tests were significant, respectively. The latter category included 19 different markers and 12 samples with a maximum of three significant tests for the same sample. Only one significant test involved a SNP likely to be affected by selection. A total of 726 865 tests for LD within samples were performed, of which 1309 tests remained significant ($P < 0.05$) after correction for multiple testing. These ranged between 29 and 95 significant outcomes (of ~40 000 possible SNP pairs) within samples, and no SNP pair was significant in more than three of 21 samples. We thus do not expect LD or departure from HWE across the 281 loci to affect downstream analyses. Levels of H_e and H_o were similar and ranged between 0.27 and 0.32 with no clear spatial pattern of regionally differentiated levels of genetic diversity (Table 1).

Outlier tests for selected vs. neutral variation

For the global analyses, overall 16 (5.7%) and 14 (5.0%) outlier loci were suggested to be under divergent selection at the 5% and 1% thresholds, respectively, across the two genome scan approaches. All 16 loci were detected by BAYESCAN, while 15 of these were also detected by ARLEQUIN (Fig. 2a, Table 2). A comparison of genotyping error across all loci (1.57%) vs. the 16 global outlier loci (1.08%) affirmed that outliers were not expected to suffer increased genotyping error rates. No outliers for balancing selection were observed (Fig. 2a). Based on the combined inference from global outlier tests, the 'full' marker set included all loci (16 outlier loci and 265 'neutral'), whereas the 'neutral' marker set included 265 putatively neutral loci. Genome scans comprising only samples across each of the two temperature clines and each of the two salinity clines

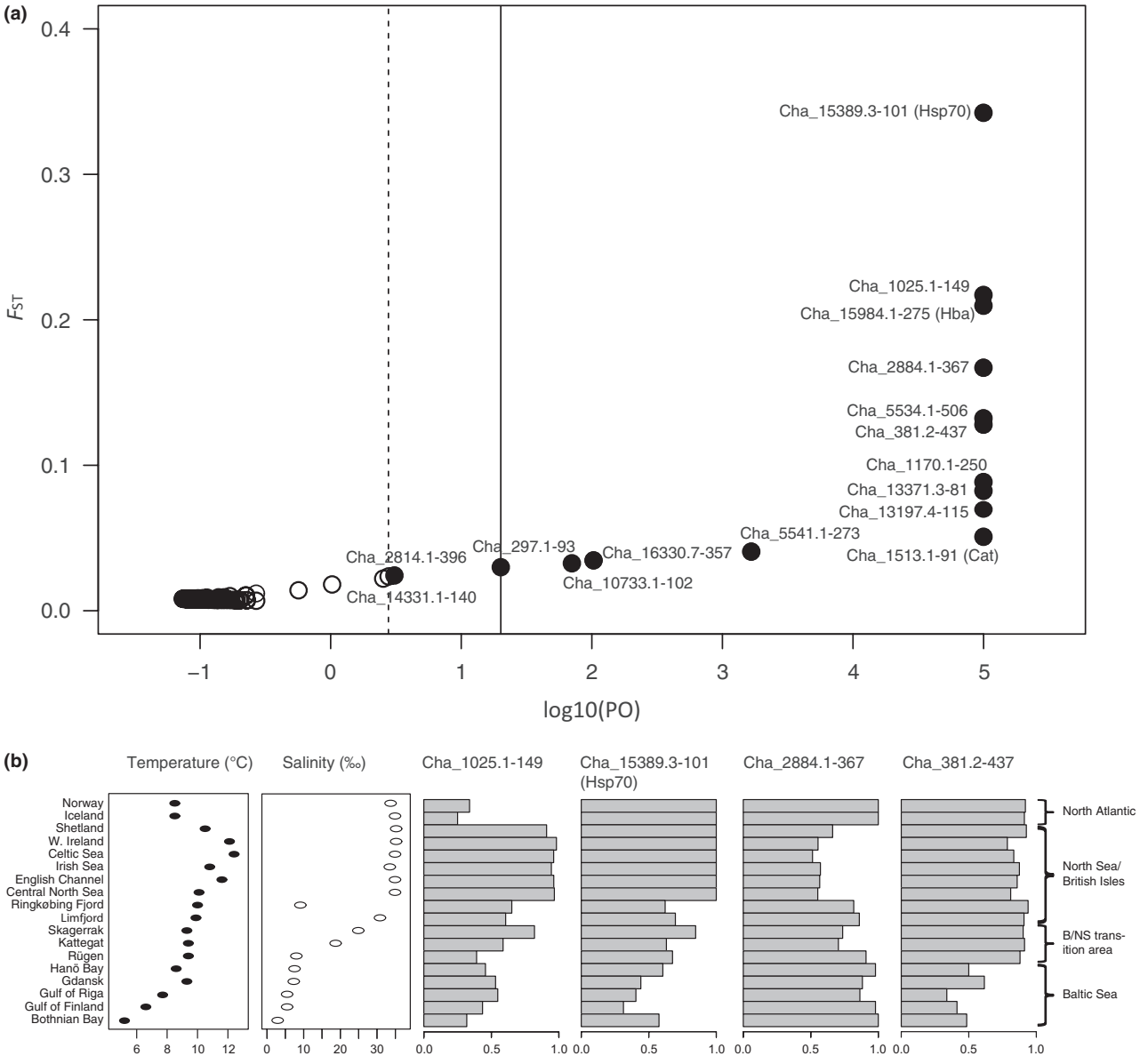


Fig. 2 Global genome scan and allele frequency plots of candidate SNPs for divergent selection. (a) Result from the global genome scan analysis using the approach by Foll & Gaggiotti (2008). Broken and solid lines represent the 5% and 1% log10(PO) thresholds for being under selection after correction with false discovery rate. Solid black circles denote SNPs that were also significant outliers using a hierarchical genome scan (Excoffier *et al.* 2009). All global outliers above the 5% level are identified by their SNP name. (b) Sample-specific values for annual average temperature, salinity and allele frequencies for a subset of four global outliers. Samples are ordered by geographical connectivity. B/NS = Baltic Sea/North Sea.

Table 2 Global results for selection including all samples

Outlier results			BAYENV results				
ARLEQUIN	BAYESCAN	SNP (Cha_)	Lat	Long	SST	SST.spawn	PSU
**	**	1025.1-149	**	**	**	(**)	(**)
*	**	10733.1-102					
**	**	1170.1-250	**		(**)		
**	**	13197.4-115		*			
**	**	13371.3-81		*	**		**
	*	14331.1-140					
**	**	1513.1-91 (Cat)		*			**
**	**	15389.3-101 (Hsp70)		**	**	(**)	**
**	**	15984.1-275 (Hba)		**	*	(**)	**
**	**	16330.7-357					
*	*	2814.1-396					
**	**	2884.1-367	**	*	**	(*)	(**)
*	**	297.1-93					
**	**	381.2-437		**	(**)		**
**	**	5534.1-506	**		**		
**	**	5541.1-273					

Overview of results from ARLEQUIN and BAYESCAN outlier tests (left of SNP names) together with all landscape association (BAYENV) results. The first two columns left of the SNP names show all detected outliers where * and ** denote outliers with $P < 0.05$ or 0.01 for the ARLEQUIN analysis. BAYESCAN outliers were detected with false discovery rates of 5% (*) and 1% (**). Statistical inference of correlations between SNPs and landscape parameters are given for relationships with log10(BF) = 1.5–2.0 (*) and log10(BF) > 2.0 (**). Corresponding correlations from partial Mantel tests that became nonsignificant when controlling for geographical distance are shown in parentheses. No relationships between neutral SNPs and tested landscape parameters had log10(BF) > 1.5 in the BAYENV tests (not shown). Lat, latitude; Long, longitude; SST, annual mean temperature, SST spawn, average spawning season temperature; PSU, annual mean salinity.

revealed between 11 and 14 outliers, with the majority only detected by ARLEQUIN (Tables 3–4). All 14 global outliers detected at the 1% thresholds were also detected in one or more regional tests. Two loci were outliers across both temperature clines (Table 3), and seven loci were outliers across both salinity clines (Table 4). In total, 39 loci were identified as outliers in one or more analyses (Tables 2–4), and of these, 28 were annotated (Helyar *et al.* 2012).

Population structure

Clustering analysis based on the 'neutral' marker set suggested a model of $K = 3$ as the statistically most likely ($\ln(K) = -145\,231 \pm 30$ SD). Most individual genotypes indicated admixture between clusters, but overall the three identified clusters corresponded with (i) the Baltic Sea, (ii) the Baltic/North Sea transition area and (iii) the North Sea/British Isles/North Atlantic (Fig. 3a). Setting $K = 4$ decreased the probability to $\ln(K) = -145\,603 \pm 43$ SD. Here, three samples from the North Sea/North Atlantic (Shetland, Norway and Iceland) showed a trend of being admixed between a North Sea/British Isles and a fourth, North Atlantic cluster (Fig. 3a). AMOVA tests for $K = 3$ and 4 revealed similar levels of variation among groups (Table 5). When using the 'full' marker set, a model with $K = 4$ was suggested as the single most likely scenario ($\ln(K) = -153\,959 \pm 43$ SD). Again, most individuals exhibited admixed genotypes, but the four clusters overall corresponded with (i) the Baltic, (ii) the Baltic/North Sea transition area, (iii) the North Sea/British Isles and (iv) the North Atlantic (Fig. 3b). The four clusters were further supported by the AMOVA revealing increased levels of variation among four groups compared to three (Table 5). For a $K = 4$ model, the two marker sets were largely similar in defining a total of four groups, and we consider this as the most likely number of groups detectable with our data. However, in comparison, the 'full' marker set was able to more clearly define the North Atlantic cluster as well as identifying admixture between the Baltic and the North Atlantic clusters in the Bothnian Bay sample (Fig. 3).

Statistical power for detecting genetic differentiation among local populations with neutral, bi-allelic markers was high (>0.89 for detecting differentiation at $F_{ST} \geq 0.005$), based on the POWSIM analysis. Genetic differentiation between three temporal within-location replicates (representing three major clusters) inferred for both the 'neutral' and 'full' marker sets was low and in all cases nonsignificant ($F_{ST} = -0.002$ to 0.005, $P > 0.05$), suggesting that the identified structure is temporally

Table 3 Regional results for selection across two latitudinal clines (reflecting temperature gradients) in the North Sea/North Atlantic (five samples) and in the Baltic Sea (five samples), respectively

North Sea/North Atlantic							
(Norway, Iceland, Shetland, Central North Sea, English Channel)							
Outlier results			BAYENV results				
ARLEQUIN	BAYESCAN	SNP (Cha_)	Lat	Long	SST	SST.spawn	PSU
**	**	1025.1-149	**		(**)		
**	**	10733.1-102					
**	**	1170.1-250	**		(*)		
**		11922.3-225					
	**	13197.4-115		*			
**	**	13371.3-81			(**)		
*		13427.1-146					
*		16060.1-279					
**	**	2884.1-367	**		(**)		
**	*	297.1-93					
**	**	462.3-102					
**	**	5534.1-506	**		**		
Baltic Sea							
(Bothnian Bay, Gulf of Finland, Gulf of Riga, Hanö Bay, Gdansk)							
Outlier results			BAYENV results				
ARLEQUIN	BAYESCAN	SNP (Cha_)	Lat	Long	SST	SST.spawn	PSU
**		10428.2-348					
**		11521.1-298					
**		12888.1-297					
*		13197.3-287					
*		13197.4-115					
*		15056.1-166					
*		15389.3-101 (Hsp70)					
**		1567.1-307					
**		15898.2-568					
**		160.1-805					
**	*	2884.1-367					
**		535.2-394					
**		5625.1-135					
**		9634.1-256					

Results are presented as in Table 2, and underlined SNPs represent replicated outliers in both transects.

stable at least within the time frame studied. Across all pairs of samples, F_{ST} estimates based on 265 neutral markers were generally low (F_{ST} = −0.002–0.012), while estimates based on all 281 SNPs were slightly higher (F_{ST} = −0.002 to 0.028; Table S2, Supporting information). Comparisons between samples from the four major population groups were generally significant for both marker sets, whereas within-group comparisons were often low (all F_{ST} < 0.008) and nonsignificant (Table S2, Supporting information). Exceptions to overall within-group homogeneity were mostly identified for the 'full' marker set and included (i) differentiation between the Gulf of Finland and Hanö Bay within the Baltic group, (ii) samples from Rügen exhibiting differentiation from the Skagerrak and Kattegat for the Baltic/North Sea transition area group, (iii) Shetland being differentiated from the Irish Sea, Celtic Sea and English Channel samples in the North Sea/British Isles group, (iv) Central North Sea and Irish Sea samples within the North Sea/British Isles group and (v) Norway being differentiated from Iceland in the North Atlantic group (Table S2, Supporting information). We cannot rule out that some loci affected by selec-

Table 4 Regional results for selection across longitudinal clines including two low-salinity environments from the North Sea/British Isles into the Baltic Sea (eight samples) and Ringkøbing Fjord (four samples), respectively.

North Sea/British Isles—Baltic Sea							
(W. Ireland, Irish Sea, Central North Sea, Skagerrak, Kattegat, Rügen, Gdansk, Gulf of Riga)							
Outlier results			BAYENV results				
ARLEQUIN	BAYESCAN	SNP (Cha_)	Lat	Long	SST	SST spawn	PSU
**	**	**1025.1-149**		**	**	**	**
*		1170.1-250					
**	**	**13371.3-81**		*	(*)		**
*		14067.1-259					
**	**	**1513.1-91 (Cat)**		*			*
**	**	**15389.3-101 (Hsp70)**		**	(**)	(**)	**
**	**	**15984.1-275 (Hba)**		**	**	**	**
**	*	16330.7-357					
**	**	2884.1-367		**		(**)	**
**	**	381.2-437					
**		3888.1-826					
*		688.1-238					
*		693.2-263					
North Sea/British Isles—Ringkøbing Fjord							
(W. Ireland, Irish Sea, Central North Sea, Ringkøbing Fjord)							
Outlier results			BAYENV results				
ARLEQUIN	BAYESCAN	SNP (Cha_)	Lat	Long	SST	SST spawn	PSU
**	**	**1025.1-149**		*			(**)
**		10733.1-102					
**		**13371.3-81**					(*)
**		**1513.1-91 (Cat)**					(*)
**	**	**15389.3-101 (Hsp70)**		**	(**)		**
**		15964.1-332					
**	**	**15984.1-275 (Hba)**		**			(**)
**		2884.1-367					
*		318.1-301					
**		5541.1-273					
*		693.2-263					
**		7456.1-168					
*		8760.1-243					

Results are presented as in Table 3, and SNPs in boldface show correlations with one or more similar landscape parameters across both clines.

tion were included in the 'neutral' marker set, because only global outliers were excluded in defining neutral markers, despite the prevalence of other loci exhibiting regional outlier status. However, here we mainly apply an intercluster comparison of neutral and selected data, and combined with the finding of mainly weak and non-significant neutral genetic differentiation within clusters (Table S2, Supporting information), we argue that the applied approach remains useful.

Environmental associations

The global test for correlations between individual loci and six landscape variables revealed significant associations with one or more landscape variables for ten of the 16 global outlier loci (Table 2). None of the 265 neutral loci were correlated with any of the tested variables. Outlier loci showed distinct allele frequency distributions potentially reflecting the effects of spatially

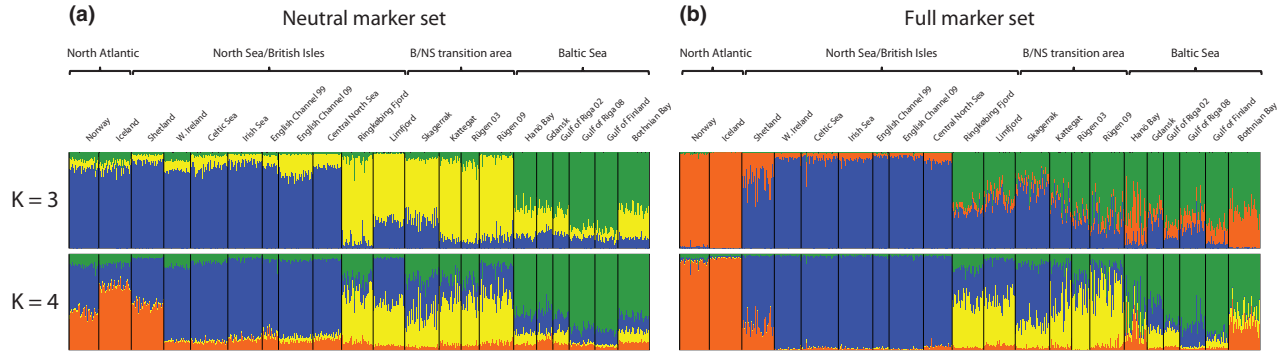


Fig. 3 Results from clustering analyses for two data sets either (a) excluding global outlier loci (neutral marker set) or (b) including all loci (full marker set) for *K* values of 3 and 4. Samples are ordered to reflect geographical connectivity illustrated by the top brackets representing the geographical origin of each sample irrespective of genetic composition. Colours designate corresponding genetic clusters between data sets: orange = North Atlantic, blue = North Sea/British Isles, yellow = Baltic/North Sea transition area and green = Baltic Sea.

Table 5 AMOVA based on neutral loci (neutral marker set) and all loci (full marker set). For each data set, an AMOVA was performed for $K = 3$ and 4 following clustering results from STRUCTURE analyses (see text for details). All variance levels are highly significant ($P < 0.001$)

		% variation	
Data set	Hierarchical level	$K = 3$	$K = 4$
Neutral marker set	Among groups (F_{CT})	0.31	0.32
	Among populations within groups (F_{SC})	0.19	0.16
	Within populations (F_{ST})	99.49	99.52
Full marker set	Among groups (F_{CT})	1.10	1.32
	Among populations within groups (F_{SC})	0.58	0.31
	Within populations (F_{ST})	98.33	98.37

varying selective forces. This is exemplified in Fig. 2b, illustrating allele frequency distributions for a subset of four global outlier loci. While allele frequencies generally vary most between the Baltic and the Atlantic, some SNPs show more similar allele frequencies between the geographically very distant North Atlantic and Baltic clusters than between geographically adjacent regions (e.g. Cha_1025.1-149 and Cha_2884.1-367). One locus (Cha_381.2-437) was mainly polymorphic in Baltic populations and was either fixed or near fixation in all other populations, whereas another locus (Cha_15389.3-101) was polymorphic throughout the Baltic and Baltic/North Sea transition area but fixed in all other populations (Fig. 2b).

All loci significantly correlated with salinity showed similar patterns with both annual and spawning period averages; thus, we only present results for annual averages below (Tables 2-4). Of seven loci globally correlated with salinity, six and five associations were also significant (log10(BF) > 1.5) across the two regional salinity clines, respectively (Fig. 4a; Table 4). Three of these loci are annotated (Helyar *et al.* 2012), including a heat-shock protein (Hsp70; Cha_15389.3-101), a haemoglobin alpha subunit gene (Hba; Cha_15984.1-275) and a gene coding for the enzyme catalase (Cat; Cha_1513.1-91).

While eight loci showed significant associations with annual mean temperature, only four of these were significantly correlated with spawning period temperature (Table 2). Likewise, none of the loci correlated with spawning period temperature across the North Sea/North Atlantic temperature cline, despite five loci showing significant associations with annual mean temperature (Fig. 4b). No landscape variables were significantly correlated with allelic variation over the Baltic temperature cline (Table 3), in spite of two outlier loci identified across both the North Sea/North Atlantic and Baltic temperature clines. Several loci that were significantly associated with temperature and/or salinity were also significantly correlated with latitude and/or longitude potentially indicating isolation-by-distance relationships covarying with environmental factors (Table 2). However, 11 of 19 global environmental correlations remained significant after controlling for geographical distance using partial Mantel tests (Table 2). Including regional tests, in total, 11 of 26 temperature correlations and 12 of 18 salinity correlations remained significant after accounting for geographical distance (Tables 2–4). Further, as the method by Coop *et al.* (2010) already controls for population demography using neutral marker information, it is not expected that isolation-by-distance effects alone explain the associations between adaptive genetic variance and environmental factors, as reported in other studies (see Vasemägi 2006).

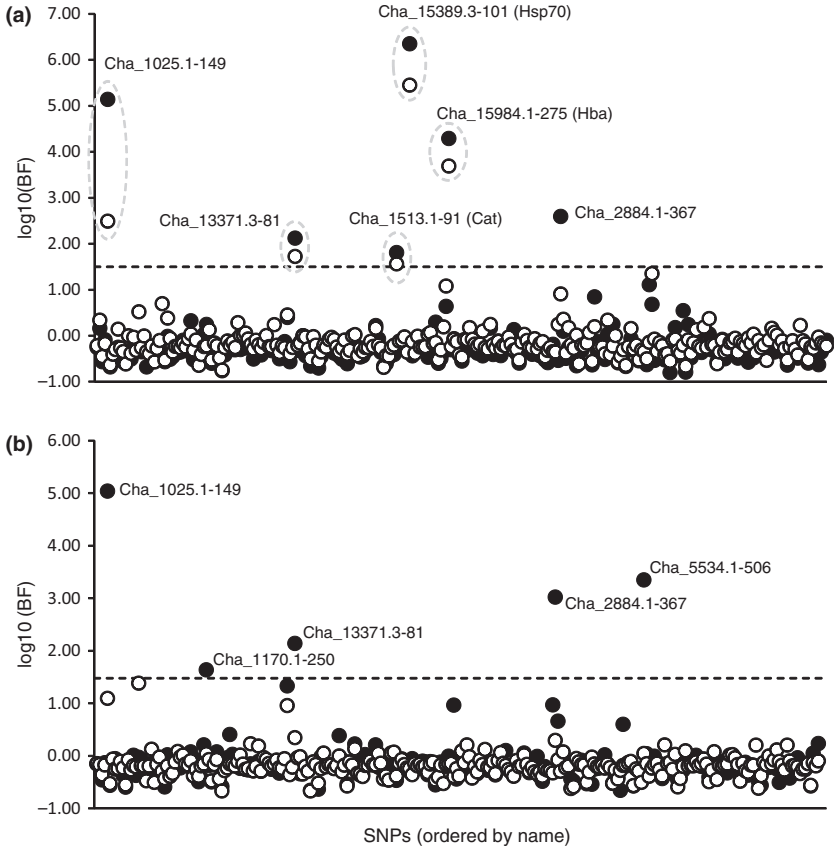


Fig. 4 Landscape association results. (a) Annual mean salinity over longitudinal clines (reflecting two low-salinity environments); 'North Sea/British Isles—Baltic Sea' (solid circles) and 'North Sea/British Isles—Ringkøbing Fjord' (open circles). See Table 4 for samples included in each analysis. Broken ellipses in grey denote five candidate loci showing significant correlations with salinity over both clines. (b) Annual (solid circles) and spawning season (open circles) mean temperatures in the North Sea/North Atlantic temperature cline. Broken horizontal lines mark lower thresholds of log10(BF) = 1.5.

Discussion

The extent and dynamics of local adaptation is key to understanding the ecological and evolutionary processes that influence biodiversity, as well as providing a spatially explicit framework for the conservation of genetic resources. While it is well recognized that opportunities for local adaptation in freshwaters due to habitat fragmentation and typically constrained dispersal (e.g. Mäkinen *et al.* 2008; Hohenlohe *et al.* 2010) are higher than in many marine fishes, recent evidence indicates remarkably small-scale adaptive variation in the latter (Moen *et al.* 2008; Poulsen *et al.* 2011). Here, by applying analyses of a large number of novel transcriptome-based SNP markers to spatio-temporal samples of herring, we identified outlier candidate genes indicating divergent selection among locally adapted populations. Moreover, candidate gene variation did not follow spatially uniform patterns across loci, suggesting that local populations have undergone multiple selective sweeps. Landscape genetic analyses suggested that environmental heterogeneity is an important driving force of divergent selection among populations, even in high gene flow organisms.

Combining neutral and selected loci to assess population structure

The highly consistent results from two statistical approaches for global outlier detection strongly suggest that the identified outlier loci or associated genomic regions are subject to divergent selection. The global outliers made up 5.7% of all analysed loci, which is within the range reported for other organisms (Nosil *et al.* 2009), including another high gene flow marine fish, Atlantic cod (Nielsen *et al.* 2009b; Bradbury *et al.* 2010). No outliers were suggested to be under balancing selection, possibly due to reduced statistical power when studying weakly structured species (Foll & Gaggiotti 2008). Substantial differences between the clustering of populations when applying the 'neutral' vs. the 'full' data set were observed. Identification of

three major clusters comprising the Baltic Sea, Baltic/North Sea transition area and the North Sea/British Isles/North Atlantic, respectively, using the 'neutral' marker set is in accordance with a previous microsatellite study (Bekkevold *et al.* 2005). An additional cluster was identified when adding outlier loci to the marker set, leading to a clear separation of North Atlantic from North Sea/British Isles populations, but also to more complex admixture patterns within populations. Such patterns of structuring clearly illustrate the increased statistical power for distinguishing weakly structured populations from the inclusion of only a few loci affected by divergent selection.

The inclusion of non-neutral markers may violate model assumptions of STRUCTURE (Pritchard *et al.* 2000) if outlier loci are under fluctuating environmental selection pressures uncoupled from the general population structuring process (migration and drift) within the species. However, differentiation at outlier loci may also elucidate evolutionary significant population units that could not be detected with neutral markers alone. Furthermore, no systematic trends of LD or deviations from HWE were observed at any of the outliers, and outliers exhibited consistent and temporally stable (over 6–10 years) patterns within and among geographical regions. Thus, we argue that careful inclusion of selected loci in a comparative approach (as presented here) remains useful for assessing spatial scales of demographically and reproductively isolated populations.

At local scales, there were clear examples of genetic separation between samples from geographically adjacent spawning locations, supported by both the clustering analysis and pairwise F_{ST} estimates for neutral markers. For example, samples from two fjords draining into the eastern North Sea (Limfjord and Ringkøbing Fjord) exhibited strong differentiation from all western North Sea locations (Fig. 3), in spite of phenotypic marker studies showing overlapping feeding habitat and large potential for mixing between populations in these areas (Rosenberg & Palmen 1982). Similarly, the western Baltic population of Rügen also exhibited clear genetic heterogeneity from the two neighbouring Baltic populations at Hanö Bay and Gdansk. In both cases, the results demonstrate that genetic variation does not follow a linear isolation-by-distance model, and corroborates natal homing as a strong driver of population structuring in herring (Gaggiotti *et al.* 2009).

Genetic divergence among samples in the North Sea/British Isles/North Atlantic was dramatically different using the 'neutral' vs. 'full' marker sets. Whereas neutral markers exhibited low, nonsignificant differentiation, agreeing with microsatellite studies (Mariani *et al.* 2005; Gaggiotti *et al.* 2009), analyses including selected loci exhibited a clear north–south separation with pairwise F_{ST} estimates of the same magnitude as between North Sea and Baltic Sea samples (Table S2, Supporting information). Such patterns could reflect selective sweeps for SNPs or associated gene regions at local scales corroborating findings in other marine fishes (Schulte 2001; Hemmer-Hansen *et al.* 2007a; Larsen *et al.* 2007; Moen *et al.* 2008; Nielsen *et al.* 2009b; Bradbury *et al.* 2010; Poulsen *et al.* 2011), and hence contribute to the notion that local selection pressures can override the homogenizing effects of high gene flow (Yeaman & Otto 2011). Indeed, results from Atlantic cod have shown similar latitudinal trends in the northeastern Atlantic Ocean to those identified for herring in this study (Nielsen *et al.* 2009b; Bradbury *et al.* 2010), and similar adaptive patterns have also been demonstrated in marine fishes inhabiting the western Atlantic (Schulte 2001; Bradbury *et al.* 2010).

The Baltic/North Sea transition area is hypothesized to constitute a hybrid zone with fish being genetically distinct from either North Sea or Baltic fish populations across several species (Nielsen *et al.* 2003; Hemmer-Hansen *et al.* 2007b; Limborg *et al.* 2009), including herring (Bekkevold *et al.* 2005). A genetically distinct cluster of herring in the Baltic/North Sea transition area was also supported here (see e.g. Fig. 3a for $K = 3$), in accordance with the findings of Gaggiotti *et al.* (2009). Further, our results suggested a relatively stronger admixture pattern for selected than for neutral loci (compare Fig. 3a and 3b). This pattern might reflect that populations in the transition area, despite exhibiting relatively closer neutral genetic relationships with Baltic than with North Sea populations for $K = 2$ (not shown), experience environmental selection pressures that are more similar to those in the North Sea. This was supported by sample-specific allele frequencies for the selected loci Cha_1025.1-149 and Cha_381.2-437, where some Baltic/North Sea transition area samples showed higher resemblance to North Sea populations (Fig. 2b).

The additional evidence provided here that adaptive divergence is marked even among potentially high gene flow species, such as herring, has wider significance. If gene flow restricts adaptive divergence, as is often assumed (Slatkin 1987), standard approaches using neutral genetic markers and landscape genetic approaches may be sufficient to get a crude estimate of adaptive variation. However, indications here, as elsewhere (Nielsen *et al.* 2009a), reinforce the notion of disparity among patterns of structuring across neutral and selected loci. Thus, if adaptive divergence does limit gene flow, genetic population structure may be poorly predicted from larval dispersal patterns, but more related to environmental heterogeneity that is sometimes obvious (Jørgensen *et al.* 2008), but sometimes not (Ha-

user & Carvalho 2008). Moreover, the implications for recruitment dynamics are considerable. Occasionally, high rates of larval influx from divergent populations may contribute little to local recruitment and may indeed be detrimental by increasing maladaptive traits (migration load). In such circumstances, selective mortality may be an important factor explaining population structure and could underlie some of the abrupt genetic discontinuities observed across hybrid zones of divergent populations, such as detected here in Baltic–North Sea herring. Documented evidence of high selective mortality in recruits to local populations (Planes & Lenfant 2002; Veliz *et al.* 2006; Vigliola *et al.* 2007) adds considerable support to this notion.

Environmental adaptation and candidate genes

Our results revealed an important role of environmental heterogeneity in shaping adaptive genetic variation at outlier genes. Specifically, the landscape genetic approach demonstrated clear associations with temperature for nine outliers and with salinity for seven outlier loci. The observation that only adaptive loci correlated with environmental factors further illustrates that divergent selection is an important force leading to locally adapted populations of herring, despite assumingly high levels of gene flow. Acknowledging the possibility that temperature or salinity is merely correlated with other environmental selection forces, our results support an evolutionary scenario with a strong environmental effect on shaping adaptive genetic variation in local herring populations.

Temperature is expected to affect a range of physiological pathways representing a multitude of underlying genes in poikilothermic organisms forced to exert innate responses to changes in ambient temperatures. Thus, it is not surprising that temperature affects a relatively large number of outlier genes including those also associated with salinity. Both temperature and salinity have also been suggested to shape adaptive genetic diversity among Atlantic cod in populations from some of the same areas as in this study (Nielsen *et al.* 2009b; Bradbury *et al.* 2010), as well as in other marine fishes (e.g. Schulte 2001; Mäkinen *et al.* 2008). However, a relatively large proportion of global outlier loci (6 of 16) did not correlate with landscape parameters, clearly suggesting an adaptive role for other (untested) selective agents such as environmental (physical, chemical, biological) factors or landscape-independent selection from intrinsic genetic incompatibilities due to, for example, epistasis (Bierne *et al.* 2011). A few outlier loci showed similar allele frequencies in the Baltic and North Atlantic samples and thus a tendency for clustering of northern Baltic and North Atlantic herring for the full marker set. These presumably adaptive signatures may reflect convergent evolution to common environmental conditions such as low temperature, which has also been shown for Atlantic cod on both sides of the Atlantic Ocean (Bradbury *et al.* 2010).

Herring are renowned for exhibiting population-specific spawning times (Cushing 1967), which might suggest local adaptation to spawning at specific seasonal temperatures. Temperature was identified as a covariant for several SNP loci in global analyses; however, these relationships were not evident at a regional scale except for the North Sea/North Atlantic latitudinal cline. This could result from reduced statistical power to detect outliers, caused by the reduced number of samples in each subanalysis, of which the method by Foll & Gaggiotti (2008) is expected to be particularly sensitive (Foll & Gaggiotti 2008). This was supported by the observation that more outlier loci were in fact detected using the approach of Excoffier *et al.* (2009). Alternatively, this observation could also be explained by increased type I and II errors and suggests that a large proportion of outliers only detected with the method of Excoffier *et al.* (2009) are in fact false positives (Narum & Hess 2011). The apparent lack of genetic covariance with temperature in the Baltic for loci exhibiting such a relationship globally contrasts with a previous microsatellite-based study (Jørgensen *et al.* 2005). This discrepancy may be attributable to the inclusion of the western Baltic Rügen population in the study of Jørgensen *et al.* (2005), in contrast to this study where it clusters with a Baltic/North Sea transition area group. Thus, weaker levels of differentiation within the Baltic proper may limit statistical power for detecting a potential relationship when excluding western Baltic samples.

Contrasting results between annual temperature and spawning temperature for the North Sea/North Atlantic temperature cline suggested that adaptation to spawning temperature was not the cause of selection at any of the loci examined here. For example, the English Channel population in the southern part of the range spawns November–January, whereas the subarctic populations spawn in April–May and August–September (Iceland) and March–May (Norway) (Table 1). As a result, Icelandic herring spawn at higher average temperatures than English Channel herring, suggesting that any temperature-related selection pressures are not specifically associated with conditions during spawning and early life stages (compared to observations in salmonids; Jensen *et al.* 2008); notably, in contrast to our findings for herring, a higher number of outlier genes correlated with spawning period (compared to annual mean) temperature in Atlantic cod (Nielsen *et al.* 2009b). These contrasting findings may reflect biological differences

between the two species. However, another important lesson learned from this study is that great caution is needed when using landscape genetic approaches, because annual estimates of environmental data may differ substantially from population-specific seasons actually affecting divergent selection. This is particularly pertinent for species exhibiting large seasonal variation in time of spawning as seen for herring. Alternatively, temperature conditions during early life stages may still impose selection at genes not associated with our marker panel. Thus, while our results support an adaptive role of temperature in general, crucial life stages and functions of outlier genes correlating with temperature remain unknown.

The clear correlation between outlier SNP variation and salinity was not driven by the Baltic populations alone, as five of six outlier loci also showed significant correlations with salinity across the Ringkøbing Fjord cline. Despite the brackish Ringkøbing Fjord population's geographical proximity to the North Sea, it has a close genetic relationship with Baltic/North Sea transition area populations likely reflecting a recent shared ancestry. Thus, we cannot rule out that the Ringkøbing Fjord population adapted to a low-saline environment as part of a larger ancestral population, with subsequent colonization of the Ringkøbing Fjord. However, whether candidate genes for salinity reflect historical adaptation in a common ancestral population, more recent parallel adaptation or a combination of the two, our results strongly indicate a general adaptive role of these genes or gene regions currently maintained in geographically isolated low-salinity environments. Three of these salinity-associated genes were annotated to known functions. A strong correlation was found between salinity and a nonsynonymous mutation (Cha_15389.3-101) in the heat-shock protein Hsp70 (Helyar *et al.* 2012), a gene family with a presumed key adaptive role in relation to environmental stress in fish (reviewed in Iwama *et al.* 1998; Basu *et al.* 2002), including European flounder (*Platichtys flesus*) (Hemmer-Hansen *et al.* 2007a; Larsen *et al.* 2008) and Atlantic cod (Nielsen *et al.* 2009b). Another outlier (Cha_15984.1-275) represented a synonymous mutation in a haemoglobin alpha subunit gene (Hba). Different variants of haemoglobin genes have been shown to be involved in local adaptation of Atlantic cod populations where different alleles are associated with divergent oxygen affinities and different temperature and hydrographical conditions (Sick 1961; Andersen *et al.* 2009). The third annotated outlier, the enzyme catalase (Cha_1513.1-91), decomposes hydrogen peroxide that is often generated at harmful levels during toxic stress responses. As such, this gene may play an important stress reaction role in marine fishes, as found for the thornfish (*Therapon jarbua*) (Nagarani *et al.* 2011). These candidate gene relationships are suggestive of environmental adaptation, albeit whether they are directly targeted by selection or exhibit hitchhiking with genomic regions of adaptive significance is not resolved.

Bierne *et al.* (2011) cautioned against interpreting significant landscape correlations as evidence for environmental adaptation at specific candidate genes. Instead, they suggested that detected outliers could represent intrinsic genetic incompatibilities uncoupled from the environment (so-called tension zones), which may become trapped in external hybrid zones driven by environmental selection. Here, candidate genes for environmental adaptation exhibited spatially distinct variation among loci, suggesting that drivers of divergence are not the same across loci and populations (Fig. 2b). Thus, we argue that it is unlikely that all outliers represent environmentally uncoupled barriers to gene flow and that a high proportion of our candidate genes indeed reflect adaptation to local environments. However, with the data at hand, we were not able to determine whether intrinsic or exogenous processes were more likely to have shaped patterns of differentiation at individual loci. To further understand the genetic architecture of fitness-related traits in these presumably locally adapted populations, studies with increasing genomic coverage (Hohenlohe *et al.* 2010; Star *et al.* 2011) and controlled rearing experiments examining genetically based fitness responses to specific environmental factors (Kawecki & Ebert 2004) are warranted.

Concluding remarks

Such local adaptation is highly relevant to fisheries management. It is not merely the conservation of genetic 'diversity' ('neutral and adaptive diversity at the DNA level') that is critical for the preservation of stocks; it is the protection of genetic 'resources' (diversity at the DNA level and its phenotypic expression at ecologically important traits). Thus, extirpation of locally adapted assemblages is of particular relevance to vulnerable species experiencing continued environmental change such as global warming or overexploitation (O'Brien *et al.* 2000). While the vulnerability of species at high trophic levels and with long generation times is widely accepted (Myers & Worm 2003), a recent study by Pinsky *et al.* (2011) showed that the majority of collapsed fisheries actually involve low-level trophic species like herring and other small pelagic fishes. Coupled with our findings, this implies that conserving the genetic 'resources' in heavily exploited species, including herring, is of paramount importance in safeguarding population resilience (Hilborn *et al.* 2003; Hauser & Carvalho 2008). Findings here constitute a basis for fur-

ther exploration of the genomic variation underlying locally adaptive traits in herring and for understanding the distribution of functionally important genetic variation in marine fishes in general.

Acknowledgements

The FishPopTrace consortium is thanked for many insightful discussions. We appreciate the assistance received from Matthieu Foll with BAYESCAN analyses and Nils Ryman for compiling a new version of POWSIM suitable for large SNP data sets. We owe many thanks to Eero Aro, Philip Coupland, Geir Dahle, Audrey Geffen, Thomas Gröhsler, Birgitta Krischansson, Ciaran O'Donnell, Henn Ojaver, Guðmundur Óskarsson, Iain Penny, Jukka Pönni, Veronique Verrez-Bagnis, Phil Watts and Miroslaw Wyszynski for providing herring samples. M.T.L. received financial support from the European Commission through the FP6 projects UNCOVER (Contract No. 022717) and RECLAIM (Contract No. 044133). The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. KBBE-212399 (FishPopTrace) and from the MariFish ERA-NET project DefineIt (ERAC-CT-2006-025989).

References

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics*, **11**, 697–709.

Andersen Ø, Wetten OF, De Rosa MC *et al.* (2009) Haemoglobin polymorphisms affect the oxygen-binding properties in Atlantic cod populations. *Proceedings of the Royal Society B-Biological Sciences*, **276**, 833–841.

Andre C, Larsson LC, Laikre L *et al.* (2011) Detecting population structure in a high gene-flow species, Atlantic herring (*Clupea harengus*): direct, simultaneous evaluation of neutral vs putatively selected loci. *Heredity*, **106**, 270–280.

Aro E (1989) A review of fish migration patterns in the Baltic. *Rapports et Procés-Verbaux Des Réunions Du Conseil International Pour l'Exploration de la Mer*, **190**, 72–96.

Basu N, Todgham AE, Ackerman PA *et al.* (2002) Heat shock protein genes and their functional significance in fish. *Gene*, **295**, 173–183.

Beaumont MA (2005) Adaptation and speciation: what can F_{ST} tell us? *Trends in Ecology & Evolution*, **20**, 435–440.

Bekkevold D, Andre C, Dahlgren TG *et al.* (2005) Environmental correlates of population differentiation in Atlantic herring. *Evolution*, **59**, 2656–2668.

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Bradbury IR, Hubert S, Higgins B *et al.* (2010) Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society B-Biological Sciences*, **277**, 3725–3734.

Conover DO, Clarke LM, Munch SB, Wagner GN (2006) Spatial and temporal scales of adaptive divergence in marine fishes and the implications for conservation. *Journal of Fish Biology*, **69**, 21–47.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

Cushing DH (1967) The grouping of herring populations. *Journal of the Marine Biological Association of the United Kingdom*, **47**, 193–208.

Dahlberg MD (1979) A review of survival rates of fish eggs and larvae in relation to impact assessments. *Marine Fisheries Review*, **41**, 1–12.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Fan JB, Oliphant A, Shen R *et al.* (2003) Highly parallel SNP genotyping. *Cold Spring Harbor Symposia on Quantitative Biology*, **68**, 69–78.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Gaggiotti OE, Bekkevold D, Jørgensen HBH *et al.* (2009) Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution*, **63**, 2939–2951.

Guillot G, Mortier F, Estoup A (2005) GENELAND: a computer package for landscape genetics. *Molecular Ecology Notes*, **5**, 712–715.

Hauser L, Carvalho GR (2008) Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish and Fisheries*, **9**, 333–362.

Hauser L, Seeb JE (2008) Advances in molecular technology and their impact on fisheries genetics. *Fish and Fisheries*, **9**, 473–486.

Helyar S, Limborg MT, Bekkevold D *et al.* (2012) SNP discovery using next generation transcriptomic sequencing in Atlantic Herring (*Clupea harengus*). *PLoS ONE, accepted pending minor revisions*.

Hemmer-Hansen J, Nielsen EE, Frydenberg J, Loeschcke V (2007a) Adaptive divergence in a high gene flow environment: Hsc70 variation in the European flounder (*Platicthys flesus L.*). *Heredity*, **99**, 592–600.

Hemmer-Hansen J, Nielsen EE, Grønkjær P, Loeschcke V (2007b) Evolutionary mechanisms shaping the genetic population structure of marine fishes; lessons from the European flounder (*Platichthys flesus* L.). *Molecular Ecology*, **16**, 3104–3118.

Hilborn R, Quinn TP, Schindler DE, Rogers DE (2003) Biocomplexity and fisheries sustainability. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6564–6568.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD Tags. *PLoS Genetics*, **6**, 23.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

Iles TD, Sinclair M (1982) Atlantic herring – stock discreteness and abundance. *Science*, **215**, 627–633.

Iwama GK, Thomas PT, Forsyth RHB, Vijayan MM (1998) Heat shock protein expression in fish. *Reviews in Fish Biology and Fisheries*, **8**, 35–56.

Jeffreys H (1961) *Theory of Probability*, 3rd edn. Oxford University Press, London, p. 432.

Jensen LF, Hansen MM, Pertoldi C *et al.* (2008) Local adaptation in brown trout early life-history traits: implications for climate change adaptability. *Proceedings of the Royal Society B-Biological Sciences*, **275**, 2859–2868.

Jørgensen HBH, Hansen MM, Bekkevold D, Ruzzante DE, Loeschcke V (2005) Marine landscapes and population genetic structure of herring (*Clupea harengus* L.) in the Baltic Sea. *Molecular Ecology*, **14**, 3219–3234.

Jørgensen HBH, Pertoldi C, Hansen MM, Ruzzante DE, Loeschcke V (2008) Genetic and environmental correlates of morphological variation in a marine fish: the case of Baltic Sea herring (*Clupea harengus*). *Canadian Journal of Fisheries and Aquatic Sciences*, **65**, 389–400.

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters*, **7**, 1225–1241.

Larmuseau MHD, Raeymaekers JAM, Ruddick KG, Van Houdt JKJ, Volckaert FAM (2009) To see in different seas: spatial variation in the rhodopsin gene of the sand goby (*Pomatoschistus minutus*). *Molecular Ecology*, **18**, 4227–4239.

Larsen PF, Nielsen EE, Williams TD *et al.* (2007) Adaptive differences in gene expression in European flounder (*Platichthys flesus*). *Molecular Ecology*, **16**, 4674–4683.

Larsen PF, Nielsen EE, Williams TD, Loeschcke V (2008) Intraspecific variation in expression of candidate genes for osmoregulation, heme biosynthesis and stress resistance suggests local adaptation in European flounder (*Platichthys flesus*). *Heredity*, **101**, 247–259.

Larsson LC, Laikre L, Palm S *et al.* (2007) Concordance of allozyme and microsatellite differentiation in a marine fish, but evidence of selection at a microsatellite locus. *Molecular Ecology*, **16**, 1135–1147.

Legendre P, Legendre L (1998) *Numerical Ecology*. Elsevier, Amsterdam.

Limborg MT, Pedersen JS, Hemmer-Hansen J, Tomkiewicz J, Bekkevold D (2009) Genetic population structure of European sprat *Sprattus sprattus*: differentiation across a steep environmental gradient in a small pelagic fish. *Marine Ecology Progress Series*, **379**, 213–224.

Mäkinen HS, Cano M, Merilä J (2008) Identifying footprints of directional and balancing selection in marine and freshwater three-spined stickleback (*Gasterosteus aculeatus*) populations. *Molecular Ecology*, **17**, 3565–3582.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**, 189–197.

Mariani S, Hutchinson WF, Hatfield EMC *et al.* (2005) North Sea herring population structure revealed by microsatellite analysis. *Marine Ecology Progress Series*, **303**, 245–257.

Maynard Smith J, Haigh J (1974) The hitchhiking effect of a favorable gene. *Genetical Research*, **23**, 23–35.

Moen T, Hayes B, Nilsen F *et al.* (2008) Identification and characterisation of novel SNP markers in Atlantic cod: evidence for directional selection. *BMC Genetics*, **9**, 18.

Myers RA, Worm B (2003) Rapid worldwide depletion of predatory fish communities. *Nature*, **423**, 280–283.

Nagarani N, Devi VJ, Kumaraguru AK (2011) Mercuric chloride induced proteotoxicity and structural destabilization in marine fish (*Therapon jarbua*). *Toxicological and Environmental Chemistry*, **93**, 296–306.

Narum SR, Hess JE (2011) Comparison of F_{ST} outlier tests for SNP loci under selection. *Molecular Ecology Resources*, **11**, 184–194.

Nielsen EE, Hansen MM, Ruzzante DE, Meldrup D, Grønkjær P (2003) Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea revealed by individual admixture analysis. *Molecular Ecology*, **12**, 1497–1508.

Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009a) Population genomics of marine fishes: identifying adaptive variation in space and time. *Molecular Ecology*, **18**, 3128–3150.

Nielsen EE, Hemmer-Hansen J, Poulsen NA *et al.* (2009b) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology*, **9**, 11.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

O'Brien CM, Fox CJ, Planque B, Casey J (2000) Climate variability and North Sea cod. *Nature*, **404**, 142.

Palumbi SR (1994) Genetic divergence, reproductive isolation, and marine speciation. *Annual Review of Ecology and Systematics*, **25**, 547–572.

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Pinsky ML, Jensen OP, Ricard D, Palumbi SR (2011) Unexpected patterns of fisheries collapse in the world's oceans. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 8317–8322.

Planes S, Lenfant P (2002) Temporal change in the genetic structure between and within cohorts of a marine fish, *Diplodus sargus*, induced by a large variance in individual reproductive success. *Molecular Ecology*, **11**, 1515–1524.

Poulsen NA, Hemmer-Hansen J, Loeschcke V, Carvalho GR, Nielsen EE (2011) Microgeographical population structure and adaptation in Atlantic cod *Gadus morhua*: spatio-temporal insights from gene-associated DNA markers. *Marine Ecology Progress Series*, **436**, 231–243.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Raymond M, Rousset F (1995) Genepop (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Rosenberg R, Palmen LE (1982) Composition of herring stocks in the Skagerrak–Kattegat and the relations of these stocks with those of the North Sea and adjacent waters. *Fisheries Research*, **1**, 83–104.

Rosenblum EB, Novembre J (2007) Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. *Journal of Heredity*, **98**, 331–336.

Ruzzante DE, Mariani S, Bekkevold D *et al.* (2006) Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. *Proceedings of the Royal Society B-Biological Sciences*, **273**, 1459–1464.

Ryman N, Palm S (2006) POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Molecular Ecology Notes*, **6**, 600–602.

Schmidt PS, Serrao EA, Pearson GA *et al.* (2008) Ecological genetics in the North Atlantic: environmental gradients and adaptation at specific loci. *Ecology*, **89**, S91–S107.

Schulte PM (2001) Environmental adaptations as windows on molecular evolution. *Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology*, **128**, 597–611.

Sick K (1961) Haemoglobin polymorphism in fishes. *Nature*, **192**, 894–896.

Slatkin M (1987) Gene flow and the geographic structure of natural-populations. *Science*, **236**, 787–792.

Star B, Nederbragt AJ, Jentoft S *et al.* (2011) The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, **477**, 207–210.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

Vasemägi A (2006) The adaptive hypothesis of clinal variation revisited: single-locus clines as a result of spatially restricted gene flow. *Genetics*, **173**, 2411–2414.

Vasemägi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology*, **14**, 3623–3642.

Veliz D, Duchesne P, Bourget E, Bernatchez L (2006) Stable genetic polymorphism in heterogeneous environments: balance between asymmetrical dispersal and selection in the acorn barnacle. *Journal of Evolutionary Biology*, **19**, 589–599.

Vigliola L, Doherty PJ, Meekan MG *et al.* (2007) Genetic identity determines risk of post- settlement mortality of a marine fish. *Ecology*, **88**, 1263–1277.

Ward RD, Woodwark M, Skibinski DOF (1994) A comparison of genetic diversity levels in marine, fresh-water, and anadromous fishes. *Journal of Fish Biology*, **44**, 213–232.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Yeaman S, Otto SP (2011) Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. *Evolution*, **65**, 2123–2129.

Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection balance. *Evolution*, **65**, 1897–1911.

This study represents a part of M.T.L's PhD thesis focusing on local adaptation in marine fishes. D.B. and E.E.N. supervised M.T.L. and share a general interest in population structure and mechanisms of local adaptation in marine fishes. S.J.H. is a population geneticist who applies molecular methods to improve assessment and conservation of genetic biodiversity within exploited species. M.I.T. and G.R.C. have research interests in fisheries genetics and the evolutionary biology and conservation genetics of aquatic organisms in general. R.O. is interested in the application of genetic tools to fisheries enforcement. All authors are part of the FishPopTrace (FPT) Consortium which aims at developing gene associated SNP assays for describing and understanding spatial patterns of population structure and local adaptations in marine fish. The FPT Consortium is also interested in using loci under selection for developing cost effective traceability tools for forensic use.

Data accessibility

SNP genotypes have been deposited under the DRYAD entry doi: 10.5061/dryad.2n763.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Names of the 310 screened SNP assays from Helyar *et al.* (2012) and information on genotyping success.

Table S2 Pairwise F_{ST} comparisons for the 'neutral' marker set (below diagonal) and for the 'full' marker set (above diagonal). Significant values ($\alpha = 0.05$) after correction for multiple tests using the FDR (Benjamini & Yekutieli 2001) are marked with an asterisk (*).

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Chapter 8

Signatures of natural selection among lineages and habitats in *Oncorhynchus mykiss*

Published in *Ecology and Evolution*

Ecology and Evolution

Open Access

Signatures of natural selection among lineages and habitats in *Oncorhynchus mykiss*

Morten T. Limborg[1,2], Scott M. Blankenship[3], Sewall F. Young[1,3], Fred M. Utter[1], Lisa W. Seeb[1], Mette H. H. Hansen[4] & James E. Seeb[1]

[1]School of Aquatic and Fishery Sciences, University of Washington, Seattle, Washington, USA
[2]National Institute of Aquatic Resources, Technical University of Denmark, Vejlsøvej 39, Silkeborg, Denmark
[3]Washington Department of Fish and Wildlife, 600 Capitol Way N. Olympia, Washington, USA
[4]Department of Molecular Biology and Genetics, Faculty of Science and Technology, Aarhus University, Denmark

Keywords
Candidate genes, Interleukin, Local adaptation, MHC, Salmonid, Steelhead

Correspondence
Morten T. Limborg, National Institute of Aquatic Resources, Technical University of Denmark, Vejlsøvej 39, Silkeborg, Denmark.
Tel: (+45) 35 88 31 05;
Fax: (+45) 35 88 31 50;
E-mail: mol@aqua.dtu.dk

Funded by a grant from the Gordon and Betty Moore Foundation to JES and LWS. MTL received financial support from the European Commission through the FP6 projects UNCOVER (Contract No. 022717) and RECLAIM (Contract No. 044133).

Received: 01 August 2011; Revised: 27 September 2011; Accepted: 27 September 2011

doi: 10.1002/ece3.59

Abstract

Recent advances in molecular interrogation techniques now allow unprecedented genomic inference about the role of adaptive genetic divergence in wild populations. We used high-throughput genotyping to screen a genome-wide panel of 276 single nucleotide polymorphisms (SNPs) for the economically and culturally important salmonid *Oncorhynchus mykiss*. Samples included 805 individuals from 11 anadromous and resident populations from the northwestern United States and British Columbia, and represented two major lineages including paired populations of each life history within single drainages of each lineage. Overall patterns of variation affirmed clear distinctions between lineages and in most instances, isolation by distance within them. Evidence for divergent selection at eight candidate loci included significant landscape correlations, particularly with temperature. High diversity of two nonsynonymous mutations within the peptide-binding region of the major histocompatibility complex (MHC) class II (DAB) gene provided signatures of balancing selection. Weak signals for potential selection between sympatric resident and anadromous populations were revealed from genome scans and allele frequency comparisons. Our results suggest an important adaptive role for immune-related functions and present a large genomic resource for future studies

Introduction

Inference of the structure and relatedness of natural populations exploded in the 1960s and 1970s with the development of molecular genetics and a deepening of our understanding of the genetic basis driving evolutionary change (e.g., Lewontin 1970). Our ability to resolve closely related populations evolved steadily through time with improvements in interrogation techniques (reviewed in Schlötterer 2004; Seeb et al. 2011a). Inference from single nucleotide polymorphisms (SNPs) during the last decade has sharpened our ability to observe differences among populations with addition of data from adaptively important loci (Anderson et al. 2005; Paschou et al. 2007; Helyar et al. 2011). The advances in studying functional genetic variation through genome scans (Storz 2005) have proven especially rewarding for studies aiming at linking phenotypic variations to a genotypic background in natural populations (Dalziel et al. 2009; Nielsen et al. 2009a).

For decades, population genetics of Pacific salmonids has attracted substantial attention from both managers and researchers due to their economic importance as well as their complex biology where broadly diverse life histories have been described (reviewed in Utter 2004). A genetic basis for a range of different life histories, including oceanic migratory patterns, has been described over the years (see Quinn 2005 and references therein). The Pacific salmonid *Oncorhynchus mykiss* (Fig. 1) has been extensively studied reflecting its charisma in both recreational fisheries and aquaculture (Wishard et al. 1984; Jantz et al. 1990). *Oncorhynchus mykiss* has naturally colonized a range of habitats across the Beringial region from Kamchatka, Russia, in the west

Figure 1. Wild rainbow trout (*Oncorhynchus mykiss*) in their natural environment (Photo by Finn Sivebæk).

to Mexico in the southeastern part of its native distribution (MacCrimmon 1971). Wild populations have furthermore been successfully introduced throughout the world (MacCrimmon 1971) making it a model species for investigating local adaptation in the wild (e.g., Rubidge and Taylor 2004; Narum et al. 2008; Pearse et al. 2009; Narum et al. 2010b). In *O. mykiss,* two North American lineages likely predating the last glacial maximum (Allendorf and Utter 1979) included populations along a broad coastal region of the Pacific Northwest and distinct from those inland (also referred to as redband trout) primarily east of the Cascade Range in the Upper Columbia and Fraser Rivers (Fig. 2; Allendorf and Utter 1979; Utter et al. 1980; McCusker et al. 2000). These two clades are hereafter referred to as the coastal and inland lineages. Two distinct life-history forms of *O. mykiss* include anadromous steelhead, having extensive oceanic migrations, and purely freshwater resident rainbow trout. However, generally higher among-region than within-region genetic variation for sympatric steelhead and rainbow trout supports a polyphyletic nature of the assumingly derived resident life history (Docker and Heath 2003; Heath et al. 2008; Pearse et al. 2009). Most studies to date remain inconclusive regarding potential molecular adaptations and selective agents maintaining this life-history variation (e.g., Docker and Heath 2003; Heath et al. 2008).

Recent work using SNPs in nonmodel organisms has allowed increased knowledge about the role of functional genetic variation (Nielsen et al. 2009a). A large majority of SNPs are neutral and provide useful information about neutral evolution and demographic inference. However, SNPs residing within, or linked to, expressed genes such as those that encode for stress or immune responses, may encode alleles subject to natural selection and add insight into adaptive evolutionary processes (Morin et al. 2004; Bouck and Vision 2007). As key effectors of the adaptive immune system and displaying an unequaled level of polymorphism for coding genes, loci of the major histocompatibility complex (MHC) have received intense attention as candidates for genes under selection (e.g., reviews in Bernatchez and Landry 2003; Piertney and Oliver 2006). In teleost fishes alone, a 2008 review reported that the available sequence information included 3559 MHC class I and class II allelic variants from 137 species (Wegner 2008). Salmonids are particularly well suited for quantifying selective pressures, because of the minimalistic genetic architecture of their MHC loci: whereas other vertebrates possess multiple duplicated loci for both MHCI and MHCII, salmonids have just one locus with classical MHCI function (UBA), and one classical locus for each subunit of the MHCII (DAA and DAB) (Hansen et al. 1999; Landry and Bernatchez 2001). Nonclassical loci have been found for both MHCI (Dijkstra et al. 2006; Miller et al. 2006) and MHCII (Harstad et al. 2008), but these are highly divergent, characterized by low levels of polymorphism and are functionally different from the antigen-presenting classical loci.

In salmonid fishes, classical MHC loci have been investigated as being subject to balancing selection within and among natural populations (e.g., Miller et al. 2001; Aguilar et al. 2004) or divergent selection (Landry and Bernatchez

Figure 2. Map of sampling locations. An approximate projection of the current divide between the inland and coastal lineages is shown by a thick broken line (From Behnke 1992).

2001; Miller et al. 2001; Gomez–Uchida et al. 2011). The type of selection inferred for these genes has been shown to depend on the spatial scale considered with a tendency toward balancing selection acting at smaller regional scales (e.g., Miller et al. 2001). In contrast, patterns of divergent selection have often been inferred among populations at larger spatial scales and those inhabiting different environments (Bernatchez and Landry 2003). However, divergent selection has also been found at fine spatial scales between ecotypes of the same lake system (Gomez–Uchida et al. 2011; McGlauflin et al. 2011) indicating that factors such as habitat type or correlated variables are important drivers of selection at these genes. Thus, MHC markers show great potential for understanding and disentangling complex patterns of adaptive processes in natural populations inhabiting varying habitats such as salmonids in the Pacific Northwest.

The purpose of this study was to screen a new genome-wide SNP resource in *O. mykiss* for signatures of local adaptation over large parts of its native distribution. Defining populations as all genetically differentiated samples, we compared diverse spawning habitats representing different environmental regimes including two paired steelhead and rainbow trout populations within the same rivers. We will refer to steelhead and rainbow trout as anadromous and resident populations, respectively, in order to generalize our findings for other fish species exhibiting migratory life-history variation, including sockeye salmon (*O. nerka*) (Taylor et al. 1996) and brown trout (*Salmo trutta*)(Elliott 1994). Using a panel of 276 SNPs designed with the intent of spacing loci as widely as possible across the genome (including newly developed markers previously unscreened in natural populations), we find strong neutral structure between the two major lineages. We probed outlier loci for signatures of adaptation based on different habitats and life histories and found strong adaptive signatures for immune-related genes.

Table 1. Sample information and summary statistics. Sample size (n), expected (H_E) and observed heterozygosity (H_O), allelic richness (A_R), and percent polymorphic SNPs are given. Statistics are given for pooled samples for locations with temporal replicates.

	Population name	ID	Year	Lineage	Life history	n	H_E	H_O	A_R	Percent polymorphic SNPs
1	Sustut River	SUST96	1996	Coastal	Anadromous	50	0.20	0.20	1.62	70%
	Sustut River	SUST97	1997	Coastal	Anadromous	45				
2	Skagit River	SKAG	2007	Coastal	Anadromous	59	0.25	0.24	1.80	89%
3	Skagit River	SKAGRES	2009	Coastal	Resident	23	0.09	0.09	1.45	55%
4	Green River	GREEN	2007	Coastal	Anadromous	35	0.22	0.22	1.73	80%
5	Sol Duc River	SOL	2009	Coastal	Anadromous	94	0.23	0.23	1.76	92%
6	Chehalis River	CHEH	2007	Coastal	Anadromous	95	0.22	0.21	1.69	82%
7	Deschutes River	DESCH	1999	Inland	Anadromous	95	0.18	0.18	1.60	75%
8	Crooked Fork Creek	CROOK99	1999	Inland	Anadromous	48	0.18	0.18	1.59	82%
	Crooked Fork Creek	CROOK01	2001	Inland	Anadromous	47				
9	Twisp River	TWISP	2008	Inland	Anadromous	81	0.19	0.18	1.65	85%
10	Twisp River	TWISPRES07	2007	Inland	Resident	25	0.20	0.18	1.71	86%
	Twisp River	TWISPRES08	2008	Inland	Resident	13				
11	Deadman Creek	DEADM97	1997	Inland	Anadromous	76	0.19	0.18	1.57	64%
	Deadman Creek	DEADM99	1999	Inland	Anadromous	19				

Materials and Methods

Sampling

We analyzed 823 individuals representing 15 collections ($n =$ 24–95) of *O. mykiss* from nine rivers throughout the Pacific Northwest of North America (Table 1; Fig. 2). Our collections include sampling of sympatric anadromous and resident fish from the Twisp River and sampling of allopatric anadromous and resident populations from the Skagit River (see Table 1). Temporal replicates of populations from four locations were also analyzed (Table 1) to increase sample sizes and assure consistency in observed spatial structure (Waples 1990). For Twisp River collections, all resident fish were collected further upstream compared to sampling of sympatric anadromous fish. Residents were typically collected during July, and fish were targeted visually by size (>180 mm), robust body shape, coloration (visible parr marks, spotting, bold color), or evidence of spawning (wounds, scale loss, fin tears, expressing milt); however residents were collected from many areas that were known spawning locations of steelhead. In contrast, the resident collection from the North Fork Cascade River in the upper Skagit River system (SKAGRES) was expected to represent an upstream resident population physically isolated from downstream anadromous populations (no upstream gene flow) due to the existence of an approximately 30-m high waterfall.

Molecular analyses and number of SNP markers

Genomic DNA was extracted from fresh fin or operculum tissue using QIAGEN DNeasy 96 tissue kits (Qiagen, Valencia, California, USA). PCR amplification and genotyping was performed in 96.96 Dynamic Arrays using the Fluidigm IFC thermal cyclers and BioMark instruments following the protocols of Seeb et al. (2009). All genotypes were scored automatically using the BioMark Genotyping Analysis software (Fluidigm, San Francisco, California, USA) and verified by two independent scorers. Any discrepancies were reassessed and either kept as a consensus or discarded. Furthermore, eight individuals from each 96 DNA sample plate (i.e., 9% of all samples) were genotyped twice for one-third of the SNPs on independent arrays to ensure reproducibility of results. We screened 276 SNPs (compiled from Aguilar and Garza 2008; Brunelli et al. 2008; Campbell et al. 2009; Sanchez et al. 2009; Stephens et al. 2009; Narum et al. 2010a; Abadia–Cardoso et al. 2011; Hansen et al. 2011 and unpublished sources listed in Table S1), of which 10 did not conform to Hardy–Weinberg equilibrium (HWE) or were suggested to be in linkage disequilibrium (LD), and these were excluded from further statistical analyses (see results, Appendix 1) leaving 266 informative SNPs. Another 21 were situated in six pairs and three triplets within the same coding gene, and are consequently very tightly linked (Appendix 2; Table S1). For subsequent statistical analyses relating to neutral population structure, 12 of these, together with eight suggested outliers ($P < 0.01$), were discarded in order to assume neutrality and independence among a set of 246 remaining markers (results, Appendix 1). Genome scans and landscape genetics analyses, which are based on individual marker information, were based on 266 SNPs including all 21 SNPs in known linkage groups (see below, Appendix 1).

Temporal stability of allele frequency distributions

Pairwise F_{ST} estimates among the 15 collections were generated in Arlequin 3.5 (Excoffier and Lischer 2010) using

10,000 permutations with P values corrected for multiple tests using the sequential Bonferroni method (k = 105) (Rice 1989). These tests revealed lower estimates between all pairs of temporally replicated collections ($F_{ST} = -0.002$–0.011) compared to all spatial comparisons ($F_{ST} = 0.013$–0.375). Assuming a conservative α value of 0.001, all temporal comparisons remained nonsignificant while spatial comparisons were all significant. Temporal replicates were pooled to optimize sample sizes for a total of 11 populations in subsequent analyses.

Conformance to HWE and nonrandom segregation of SNPs

Conformance to HWE was tested independently for each locus in each of the 11 populations using the MC algorithm implemented in Genepop 4.0 (Rousset 2008). P values were corrected using the sequential Bonferroni method (k = 3069) (Rice 1989). Linkage was only known for those nine groups of SNPs that were ascertained in single Sanger sequencing reads (Hansen et al. 2011; Table S1). Thus, we simply tested for nonrandom segregation of all pairs of loci within each population using Fisher's tests for gametic LD as implemented in Genepop 4.0 (Rousset 2008). Due to the high number of tests performed (i.e., 36,315 for each population), no correction for multiple tests was performed since this approach would be overly conservative and likely underestimate truly significant relationships. We followed a hierarchical approach with the following criteria for assessing LD among markers: (1) only SNPs with a minor allele frequency (MAF) > 0.10 were considered due to an expectedly high number of false positives associated with low levels of variation, (2) only locus pairs showing more than 50% significant tests ($P < 0.05$) with at least six performed tests among 11 populations, (3) for all locus pairs showing significant LD, SNPs potentially involved in multiple pairs were discarded.

Summary statistics

Individual global F_{ST} values were estimated for each locus in Genepop 4.0 (Rousset 2008) as well as over all loci. Mean expected (H_E) and observed (H_O) heterozygosity were calculated for each locus and population using GenAlEx 6.4 (Peakall and Smouse 2006). Allelic richness (A_R), a measure of the number of alleles corrected for minimum sample size, was calculated for all populations using FSTAT v2.9.4 (Goudet 1995). GenAlEx 6.4 was used to report the percentage of polymorphic loci in each population.

Spatial population structure and diversity

We used 246 neutrally behaving and individually segregating SNPs to recalculate pairwise F_{ST} estimates among all 11 populations in Arlequin 3.5 (Excoffier and Lischer 2010) using 10,000 permutations. We then used pairwise F_{ST} values (all $P < 0.001$) to generate a multi dimensional scaling (MDS) plot in ViSta 5.6.3 (Young 1996) for visualizing neutral population structure.

Signatures of selection

To detect genomic regions under selection, we used a total of 266 SNPs including 21 SNPs from known groups of tightly linked SNPs. Inclusion of known linked SNPs is expected to increase the chance of finding signatures of selection, as even closely linked SNPs may differ substantially in their level of differentiation (Gomez–Uchida et al. 2011). To test for potential related bias, analyses were repeated with a reduced dataset not including linked SNPs. First, we performed a global genome scan for nine populations (i.e., omitting the two resident populations) using the model by Excoffier et al. (2009) as implemented in Arlequin 3.5 (Excoffier and Lischer 2010). This approach simulates a neutral distribution of F_{ST} (or F_{CT}) in relation to observed heterozygosity. Observed values by locus were then projected onto this distribution, and loci lying above or below the simulated 99% confidence threshold for neutral variation were considered as candidates for divergent or balancing selection, respectively. We applied the hierarchical test by grouping populations into two groups representing the coastal and inland lineages (Table 1). We assumed a model of 10 simulated groups with 100 demes and performed 100,000 simulations. To further understand the spatial pattern of potential selection for all outliers ($P < 0.01$) detected for both F_{ST} (i.e., among all populations) and F_{CT} (i.e., between lineages here), we plotted major allele frequencies over all populations. We also performed a genome scan over all anadromous populations using BayeScan 1.0 (Foll and Gaggiotti 2008). We ran 10 pilot runs of 5000 iterations with an additional burn-in of 50,000 iterations and a thinning interval of 50 followed by a final sample size of 10,000. Results from the two genome scan approaches were compared and dual outliers were considered as strong candidates for diversifying selection.

To investigate divergent selection between migratory life-history types, we performed individual genome scans for each of the two within-river anadromous and resident population pairs (i.e., SKAG and SKAGRES as well as TWISP and TWISPRES) using 10 simulated demes and 100,000 simulations in Arlequin 3.5. Again, we plotted major allele frequencies over all populations for all outliers above the 99% confidence levels. BayeScan 1.0 was not considered here, since it is expected to perform poorly with few samples (Foll and Gaggiotti 2008).

Environmental effects on adaptive variation

We further tested for associations between landscape variables and allelic distributions for each SNP to reinforce evidence of natural selection acting on outlier loci (as opposed to false positives) (e.g., Fraser et al. 2011). Underlying

correlations between allele frequencies and landscape parameters may occur by chance due to either isolation by distance or to more similar landscapes between neighbor populations. If not taken into account, such neutral background noise is expected to lead to an increased false-positive rate (Coop et al. 2010). We applied the Bayesian linear model implemented in the software Bayenv (Coop et al. 2010) to correct this. This method uses a covariance matrix based on neutral markers to filter out signals from neutral population structure while testing for significant relationships between landscape variables and locus-specific allele distributions. Results are given for 266 SNPs (as described above), and each landscape variable as a Bayes factor (BF). This BF reflects the ratio of the posterior support given to a model where the landscape variable has a significant effect on allele distributions over an alternative model where there is no effect on the SNP. First, we estimated a covariance matrix using the 246 independently and neutrally behaving SNPs (see above). Then, we tested for correlations between each of 266 SNPs and the following variables: (1) precipitation, (2) maximum temperature, (3) minimum temperature, (4) elevation, (5) latitude, and (6) longitude (Appendix 3). We used multiple independent Markov chain Monte Carlo (MCMC) runs with chain lengths of 100,000 iterations to ensure convergence of the model.

Genetic variation at MHC genes

Some of the loci that we used were previously annotated and represent potential candidate genomic regions for selection (Table S1). In this study, we restrict our *a priori* focus to six newly developed markers residing within the classical MHC class I (Omy_UBA3a, Omy_UBA3b, and Omy_UBA2a) and the classical MHC class II (Omy_DAB-431, Omy_DABb, and Omy_DABc) genes (Hansen et al. 2011; Table S1). We observed intriguingly high diversity at two of the MHC class II SNPs (Omy_DAB-431, Omy_DABb) known to be nonsynonymous (Hansen et al. 2011). To test for balancing selection on this gene, we reconstructed most likely haplotypes from the three SNPs in known linkage within the MHC class II gene (Table S1) using the program PHASE V2.1 by Stephens et al. (2001). We then used the Ewens–Watterson homozygosity test as implemented in Arlequin 3.5 (Excoffier and Lischer 2010) to test for balancing selection on reconstructed haplotypes within populations and overall assuming an infinite allele mutation model and using 10,000 simulations. This test compares the expected HW homozygosity based on observed haplotype frequencies (here designated as Observed *F* value) with a simulated value (Expected *F* value) expected at mutation drift equilibrium for a gene with a similar number of alleles (Ewens 1972; Watterson 1978), and where balancing selection will lead to smaller observed than expected *F* values.

Results

Laboratory analyses and tests for HWE and LD

We excluded 18 individuals with missing data at more than 50% of the loci, which likely reflected poor DNA quality. Six of the 276 SNPs that we initially screened showed significant deviation from HWE in five or more populations after correcting for multiple tests and were excluded from further analyses (Appendix 1; Table S1). We observed seven pairs of loci showing significant LD ($P < 0.05$) in more than half of the performed tests leading to the exclusion of four loci, of which some were involved in multiple pairs, to avoid potential pseudoreplication by including markers in LD (Table S1). Further analyses were based on a final dataset of 805 individuals representing 11 populations ($n = 23$–95) and 266 SNPs (Appendix 1).

Summary statistics

Over all populations, these 266 SNPs showed varying levels of differentiation with global locus-specific F_{ST} values ranging from 0.00 to 0.68. The frequency of polymorphic loci varied from 55 to 92% among populations (Table 1). We observe intermediate levels of genetic diversity ranging from 0.09 to 0.25 (H_E), 0.09 to 0.24 (H_O), and 1.45 to 1.80 (A_R) with highly reduced levels in the SKAGRES population (Table 1).

Spatial population structure and diversity

Spatial population structure inferred from significant pairwise F_{ST} values supported a pattern where most of the variation is likely caused by genetic drift and limited gene flow from historical isolation of the two lineages (Fig. 3). However, because most of the variation plotted for DIM 1 in the MDS plot was driven by the isolated SKAGRES population (Fig. 3A), omitting this population improved resolution for inferring spatial structure among remaining populations (Fig. 3B). The five coastal lineage populations cluster according to geography where the distant SUST population separates from a Puget Sound group (SKAG and GREEN) and a coastal Washington group (SOL and CHEH). Observed population structure within the inland lineage reflects contemporary geographic isolation with clear differentiation of the DEADM population (Canada) from remaining populations within the Columbia River drainage (Fig. 3B).

Signatures of selection

Candidates for balancing selection could not be distinguished from loci with observed F_{ST} values of zero, and we did not consider these further. The global genome scan assuming a hierarchical island model in Arlequin 3.5 revealed eight significant outliers for divergent selection ($P < 0.01$) at the F_{ST} level (Fig. 4A) of which five loci were also candidates at the

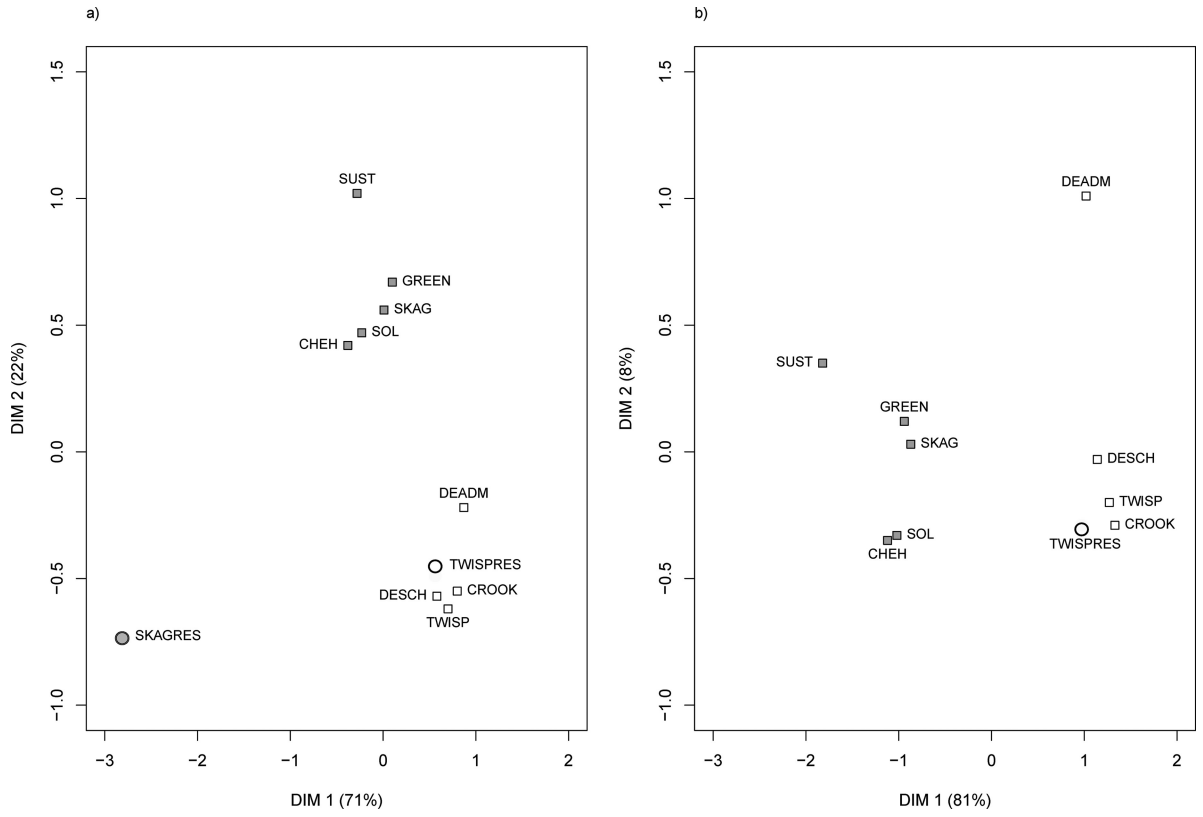


Figure 3. Multi dimensional scaling (MDS) plot showing: (a) spatial population structure for all populations including the two resident populations, and (b) a similar plot without the SKAGRES population. Population symbols follow Figure 2.

F_{CT} level (Fig. 4B). This finding is more than twice as many as expected by chance alone (1% of 266 = 2.7). BayeScan detected four of the eight global outliers ($P < 0.01$) found with Arlequin as well as three new outliers (Table S1). These three outliers only detected by BayeScan were all characterized by very low minor allele frequencies (0.001–0.012), leaving them essentially uninformative. These were not considered, and we only interpret the eight candidates detected by Arlequin further. Functional roles could be inferred for two of the outliers, Omy_IL1b-163 and Omy_ndk-152, which reside within interleukin and nucleoside diphosphate kinase genes, respectively (Table S1; references therein). As expected, the five F_{CT} outliers reflect substantial differences between the two lineages (Fig. 5A), while all three outliers only detected at the F_{ST} level suggest a pattern of divergent selection within the coastal lineage as observed from deviating allele frequencies in the northern SUST compared to the other coastal lineage populations (Fig. 5B).

Outliers under potential divergent selection ($P < 0.01$) were also detected in local genome scans for selection between within river populations exhibiting alternate life histories. We observe one outlier in the allopatric Skagit River and six in the sympatric Twisp River comparisons (Fig. 4C and 4D). Allele frequency plots for these local outliers all reveal a potential effect of anadromy. For example, allele frequencies for the anadromous Twisp River population are generally more similar to other inland anadromous populations compared to the sympatric resident population (Fig. 6A). For the outlier detected from the Skagit river populations, this pattern is even more pronounced (Fig. 6B). When also considering outliers at the 95% level, SNPs within interleukin genes (Omy_IL17–185 and Omy_IL6–320) are observed as outliers in the Skagit and Twisp River comparisons, respectively (Fig. 4C and 4D; Table S1). Another marker (Omy_97954–618) appears as an outlier for both locations at the 95% level suggestive of a consistent pattern of diversifying selection (Table S1).

Environmental correlates

Bayesian inference for correlation between locus-specific allele distributions and landscape variables showed that 12 of 17 (71%) global candidates for divergent selection ($P < 0.05$; Table S1) significantly correlated with one or more variables (Table 2), contrasted with only 11 of 246 neutrally behaving loci (4%). When only considering "decisive" relationships (i.e., log10 (BF) > 2), only outlier loci correlated with any of

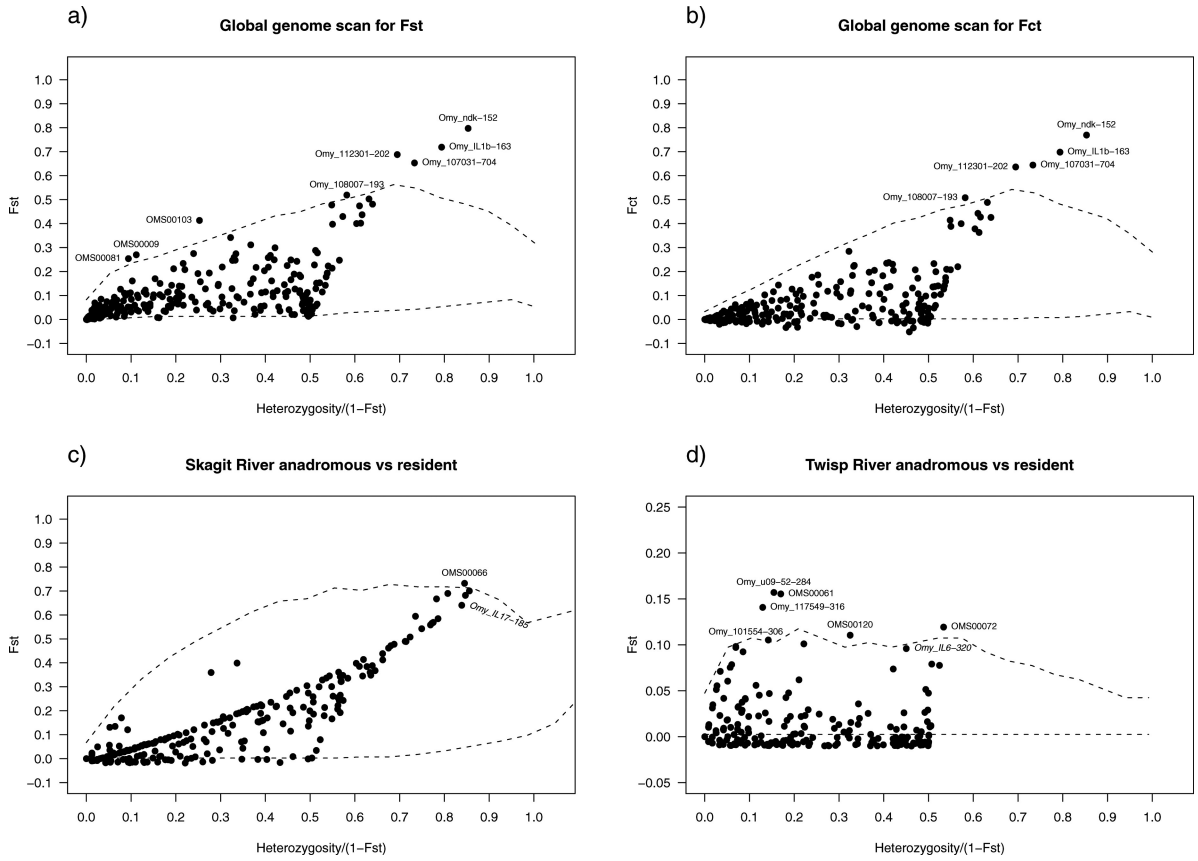


Figure 4. Outlier tests for identifying signatures of selection. (a) F_{ST}-based global test assuming hierarchical structure by grouping all anadromous populations within each lineage into two major groups. (b) F_{CT}-based global test assuming hierarchical structure as in (a). Local outlier tests for the Skagit River (c) and Twisp River (d) anadromous and resident population pairs are also shown. All outliers above the 99% confidence threshold are labeled including two interleukin genes at the 95% threshold (c and d) shown in italic. Plotted heterozygosity values are scaled by estimates of within population heterozygosity (h_0) and locus specific F_{ST} as: ($H_1 = h_0/[1 - F_{ST}]$) as described in Excoffier et al. (2009).

the variables (Table 2). Particularly precipitation and temperature appear promising for explaining patterns of divergent selection at some of the candidate loci or linked genomic regions found here.

Genetic variation at MHC genes

One SNP within the MHC class I gene (Omy_UBA2a) was discarded due to significant deviation from HWE (Table S1). Remaining MHC markers conformed to neutrality in genome scans coupled with low diversity in three of five markers (Appendix 4). However, the two nonsynonymous mutations residing within the MHC class II gene (Omy_DAB-431 and Omy_DABb) exhibit high levels of variation throughout most populations (Appendix 4). Reconstructed haplotypes based on three SNPs within the MHC class II gene revealed a significant deviation from neutrality toward balancing selection for two of 11 populations (Table 3). Furthermore, three populations had *P* values below 0.10 and all populations, except the Skagit River resident (SKAGRES), had smaller than expected *F* values pointing toward balancing selection (Table 3).

Discussion

We found highly significant spatial structure with increased levels of neutral differentiation between and within the two major lineages. This result is consistent with previous studies on *O. mykiss* from this region using allozymes and mtDNA (e.g., Allendorf and Utter 1979; McCusker et al. 2000). Applying multiple independent analytical steps (e.g., genome scans, landscape genomics, and raw allele frequency plots), the accumulated evidence supports local adaptation at several genomic regions including immune response genes.

Spatial population structure and diversity

Most of the presumed neutral genetic variation that we observed can be explained by a model of historical vicariance.

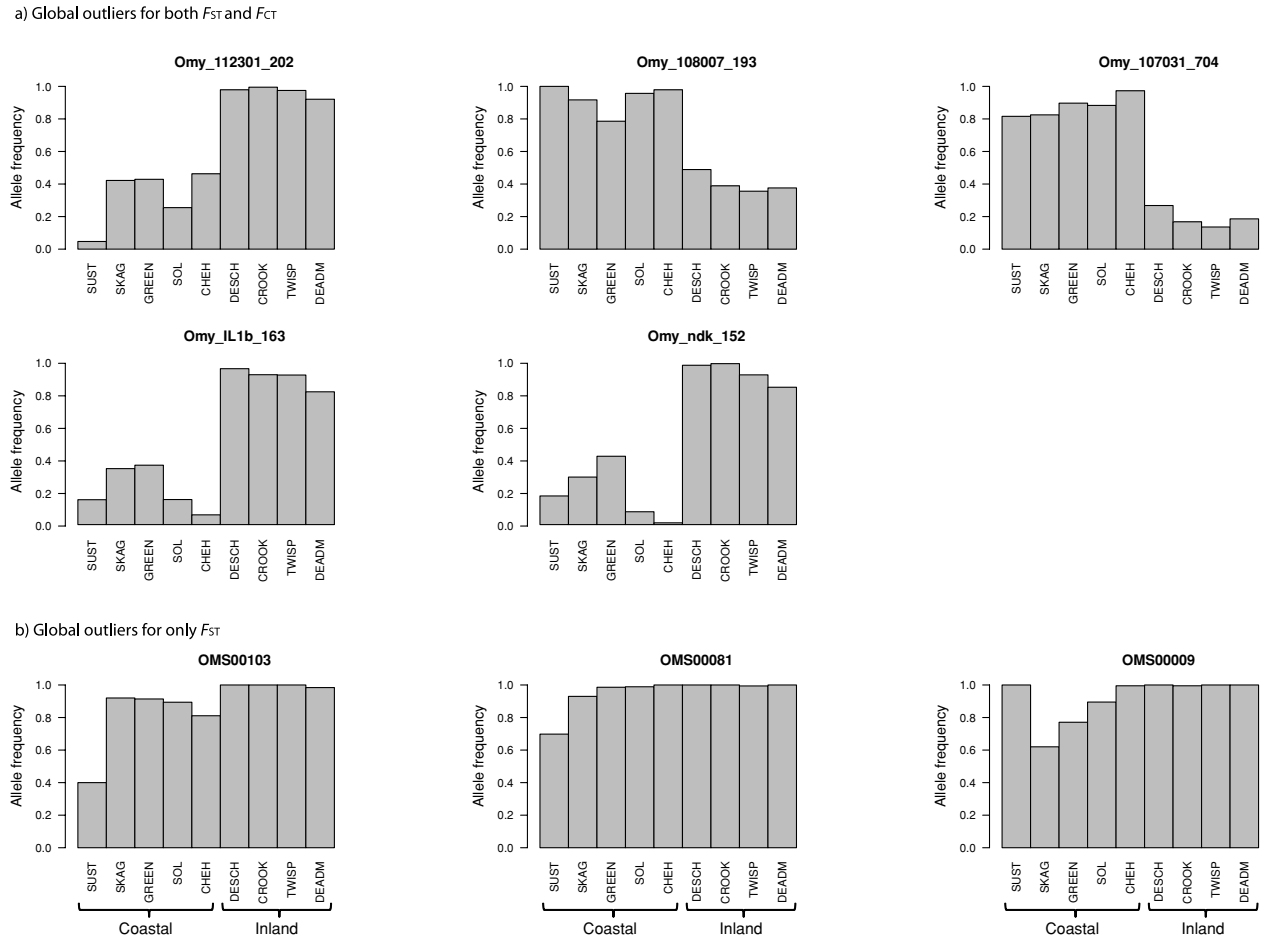


Figure 5. Frequency plots of major allele frequencies for loci detected as outliers ($P < 0.01$) in the global genome scan including nine anadromous populations. (a) Allele frequencies for five outliers detected at both the F_{ST} and F_{CT} level. (b) Allele frequencies for three outliers only detected at the F_{ST} level.

Distinct evolutionary lineages have seemingly accumulated genetic differentiation through genetic drift over glacial periods. Contemporary gene flow and drift appear less important at this large spatial scale but probably play a greater role at smaller regional scales among more recently diverged populations. This observation is supported by previous phylogenetic observations (Bagley and Gall 1998; McCusker et al. 2000) and early allozymes studies (Utter et al. 1980) showing similar patterns of strong differentiation between inland and coastal lineages compared to population structure within lineages. The location above a waterfall of the SKAGRES population likely explains its divergence (e.g., Fig. 3A) and low genetic diversity (Table 1) as a reflection of limited gene flow and low effective population size (N_e) with consequently strong genetic drift. A similar scenario has been shown for another physically isolated population of resident *O. mykiss* (Pearse et al. 2009; Martínez et al. 2011). Omitting the SKAGRES population revealed further regional structure within the coastal lineage, suggesting more recent population histories potentially coupled with higher contemporary gene flow among populations within the Puget Sound and western Washington coastal regions, respectively (Fig. 3B). Despite the large spatial scale, these observations hold great promise for applying this SNP panel in future studies focusing on smaller geographic scales. Variation between the geographically distant populations from Sustut River and Deadman Creek in BC, Canada (Fig. 3B) is also apparent. The increased differentiation observed for the Canadian populations within respective lineages is likely a reflection of the substantial geographical separation coupled with longer divergence from the other populations. It is noteworthy how weak this latter signal is compared to that observed between lineages, suggesting a limited reflection of contemporary genetic drift compared to signals from historical separation.

Despite potential gene flow between the sympatric anadromous and resident Twisp River populations (see also Christie

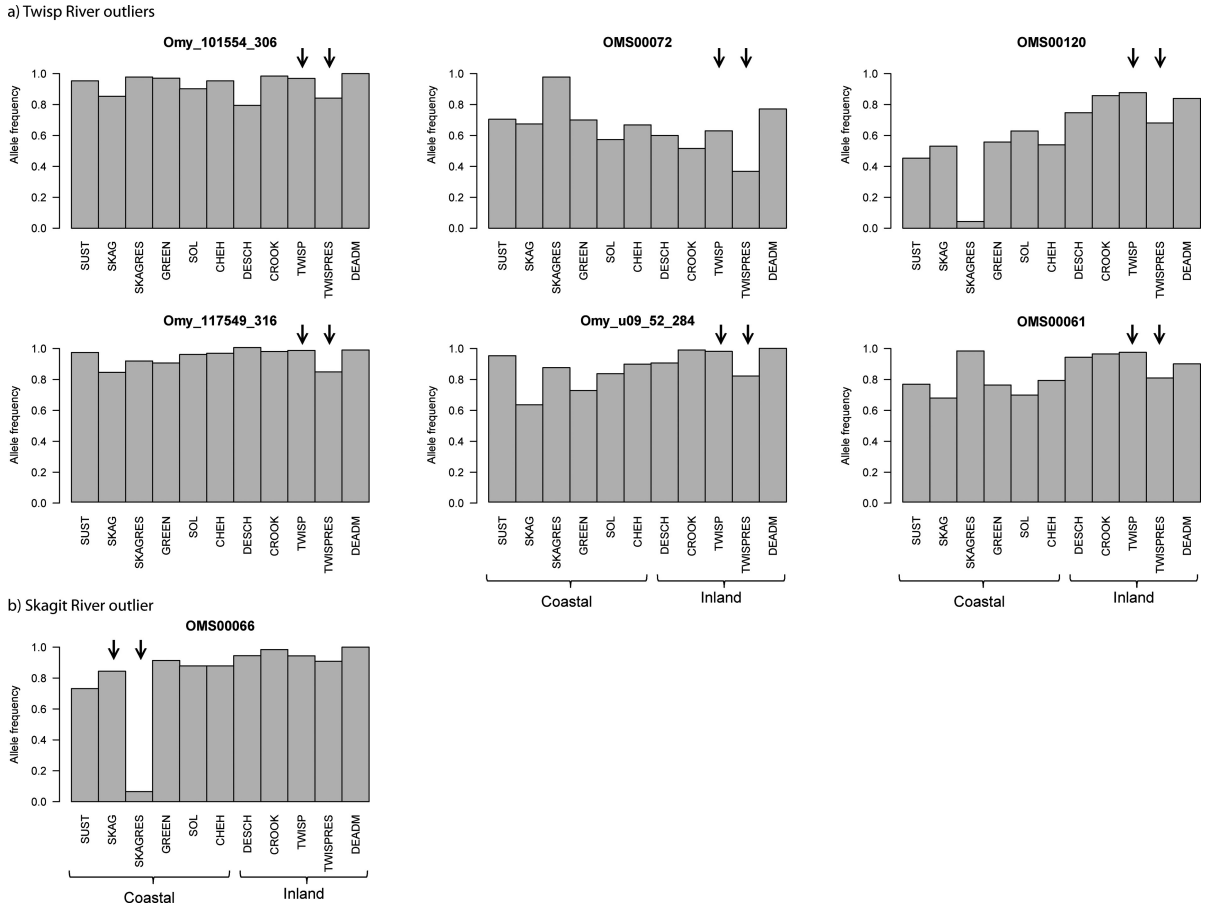


Figure 6. Frequency plots of major allele frequencies for loci detected as outliers ($P < 0.01$) between sympatric and resident population pairs. (a) Allele frequencies for six markers detected as outliers between the Twisp River populations. (b) Allele frequencies for one outlier detected between the populations within the Skagit River. Arrows denote populations included in the local genome scans.

et al. 2011), our results suggest some level of neutral population structure (pairwise $F_{ST} = 0.01$; $P < 0.001$). Zimmerman and Reeves (2000) found evidence for reproductive isolation of sympatric steelhead and rainbow trout in the Deschutes River, Oregon, and explained this with variation in timing and location of spawning activities. A similar scenario with some level of spatio-temporal reproductive isolation of the two Twisp River populations would be in accordance with our observations.

Signatures of spatially divergent selection

We detected eight candidates for directional selection among all populations (Fig. 4A). Five of these were in accordance with divergent selection between the coastal and inland lineages (Figs. 4B and 5A). Allele frequency plots (Fig. 5A) reveal high levels of information from the five F_{CT} outliers for distinguishing between the two lineages. Population history and dynamics of populations spawning within or in close proximity to the transition zone has been difficult to infer in previous studies applying fewer and assumingly neutral markers (Currens et al. 2009; Blankenship et al. 2011). Outliers observed in our study appear promising for future investigations of the nature (e.g., distinguishing neutral and adaptive genetic variation) and extent of this transition zone in *O. mykiss*.

Two outliers were known to reside within an interleukin gene (Omy_IL1b-163) and a nucleoside diphosphate kinase (Omy_ndk-152) gene (Table S1). A recent study also found Omy_ndk-152 and another SNP within an interleukin gene to be affected by selection in relation to anadromy at a much finer scale among populations within the Klickitat River (all derived from the inland lineage) draining into the Columbia River system (Narum et al. 2011). Our broad spatial representation of anadromous populations limits the ability to identify the geographic scale at which selection acts upon

Table 2. Results from Bayesian inference of locus-specific landscape correlations. Gray cells denote a locus–parameter relationship with a log10 (BF) between 1.3 and 2.0, which can be interpreted as a *P*-value between 0.01 and 0.05. Black cells represent decisive relationships with log10 (BF) > 2.0 or equivalent *P*-values below 0.01. Here, global outliers include loci at the 5% significance level (see Table S1, Supporting information).

		Tested variables*					
SNP	Selection	Precip	Tmax	Tmin	Elev	Lat	Long
OMS00180	Outlier						
OMS00118	Outlier						
Omy_star-206	Outlier						
Omy_121713-115	Outlier						
OMS00013	Outlier						
Omy_107031-704	Outlier						
Omy_112301-202	Outlier						
Omy_IL1b-163	Outlier						
Omy_ndk-152	Outlier						
Omy_108007-193	Outlier						
OMS00103	Outlier						
OMS00081	Outlier						
Omy1004	Neutral						
OMS00081	Neutral						
OMS00103	Neutral						
Omy_111005-159	Neutral						
Omy_DABb	Neutral						
Omy_u09-56-073	Neutral						
Omy_hsp47-86	Neutral						
Omy_gluR-79	Neutral						
OMS00053	Neutral						
Omy_rapd-167	Neutral						
Omy_09AAD-076	Neutral						

*Precip = annual mean precipitation (mm); Tmax = annual mean maximum temperature (°C); Tmin = annual mean minimum temperature (°C); Elev = elevation (m); Lat = latitude; Long = longitude.

Table 3. For each population, number of reconstructed haplotypes, observed, and expected levels of homozygosity (*F* value) are given with results from the Ewens–Watterson homozygosity test for deviation from neutrality at an MHC class II gene (see text for more details). *P*-values below 0.05 are highlighted in bold, and *P*-values between 0.05 and 0.10 are shown in italic.

Population	No. of haplotypes	Observed *F* value	Expected *F* value	*P*-value
CHEH	6	0.289	0.467	0.104
CROOK	6	0.295	0.468	0.113
DEADM	5	0.317	0.527	*0.076*
DESCH	5	0.255	0.530	**0.009**
GREEN	5	0.328	0.463	0.187
SKAG	5	0.313	0.500	*0.096*
SKAGRES	2	0.841	0.777	0.569
SOL	6	0.277	0.467	*0.079*
SUST	4	0.500	0.606	0.363
TWISP	6	0.248	0.457	**0.034**
TWISPRES	6	0.312	0.409	0.274
Mean	5.1	0.361	0.516	0.173
SD	1.2	0.173	0.101	0.168

outlier loci. However, by applying a hierarchical island model and comparing outliers at the F_{ST} and F_{CT} levels, we can deduce whether divergent selective forces are likely to dominate within or between the two lineages (Fig. 5). A recent study by Meier et al. (2011) showed that both number and types of outlier loci for divergent selection varied substantially at different spatial scales in brown trout. This pattern demonstrates the need for denser sampling of populations if the goal is to increase the spatial resolution of inferred selective processes. The observed outliers might therefore be shaped by heterogeneous landscapes, or other selective agents, operating at smaller geographic scales within each lineage (e.g., Narum et al. 2008; Narum et al. 2010b). Indeed, our landscape genomics analysis suggested an important link between landscape variables and several loci (Table 2). Due to the inherent uncertainty of correlations between predefined variables such as used in this study, we refrain from concluding direct functional relationships for specific loci or landscape parameters (see also Bierne et al. 2011). Nevertheless, looking at overall trends two main findings can be inferred from this analysis. First, genetic variation associated with surrounding landscape variables was dominated by outlier loci suggesting a general pattern of local adaptation to specific environments by *O. mykiss*. Second, precipitation and temperature (or correlated factors), in particular, may play important roles in shaping adaptive genetic variation in *O. mykiss*. An effect of temperature would be in accordance with three regional (F_{ST}) outliers following a latitudinal trend within the coastal lineage (Fig. 5B). A recent study by Wenger et al. (2011) also suggests an important role of temperature and flow regime (expected to be partly correlated with precipitation) in determining the distribution of suitable habitat for *O. mykiss*, adding support for crucial adaptive roles of these environmental parameters. However, these association-based findings remain indirect in nature, but direct links between variations in genotype, phenotype, and fitness remain very rare for any organism. Future studies obtaining much denser genomic coverage (see e.g., Hohenlohe et al. 2010) will allow a more direct chromosomal location of the gene(s) actually under selection. Alternatively, surveillance of fully controlled populations allows to track gene frequencies over time after being exposed to a new environment (e.g., Barrett et al. 2008). Thus, our results can be seen as hypothesis generating for future studies specifically investigating effects of certain landscape variables or candidate genes.

Migratory life-history types

We identified one outlier potentially under divergent selection between Skagit River resident and anadromous populations (Fig. 4C). For the Twisp River resident and anadromous populations, we found six putative outliers for

divergent selection (Fig. 4D). The one outlier observed between the two Skagit River populations is consistent with that expected by chance alone. Furthermore, true outliers can be difficult to distinguish from false positives in this comparison considering the high levels of observed neutral differentiation between these two populations (pairwise F_{ST} = 0.30, $P < 0.001$). However, the observation of six outlier loci (i.e., 2.3%) at the 99% confidence level between the Twisp populations, together with another marker showing signatures of selection in both locations, suggest ongoing selection between anadromous and resident life-history types. This overall result is consistent with recent studies identifying signatures of divergent selection between different migratory variants in *O. mykiss* (Martínez et al. 2011; Narum et al. 2011). All outlier loci show allelic patterns suggestive of distinctions between resident populations and the anadromous counterparts within the same lineages. For example, allele frequency plots reveal a consistent pattern of higher similarity among all inland anadromous populations than between the anadromous and resident Twisp River populations (Fig. 6A). These differences are small and at best weak indicators of ongoing selection between life histories. However, a mere effect of increased drift in an assumingly smaller and more isolated resident TWISPRES population cannot explain these observations since a general trend of increased variation was observed for this population (Table 1; Fig. 6A). Despite this generally inconclusive pattern, these outliers may potentially prove rewarding in future studies with a more targeted focus on studying selection between these life histories.

Evidence of selection acting on immune response genes

The two known nonsynonymous mutations in the peptide-binding region of a MHC class II gene (Omy_DAB_431 and Omy_DABb) generally showed high levels of diversity in most populations with MAF ranging between 0.26 and 0.48 (Appendix 4). Although only two populations gave significant results in direct tests for balancing selection acting on reconstructed haplotypes of the MHC class II gene, a clear overall trend toward balancing selection was revealed (Table 3). Lack of more significant findings may be due to limited statistical power from the limited number of alleles (Table 3) when reconstructing haplotypes from only three segregating SNPs. However, since these mutations change amino acids in the crucial peptide-binding region of the MHC class II molecules, we would expect that lack of balancing selection would have led to elimination of otherwise assumingly deleterious mutations in just a few generations. While our results are only indicative of balancing selection within or among populations, many previous studies have detected patterns of balancing selection acting on MHC loci in other salmonid fishes (Landry and Bernatchez 2001; Miller et al. 2001), including *O. mykiss* (Aguilar and Garza 2006). Furthermore, a recent study by Martínez et al. (2011) detected divergent selection on a microsatellite locus linked to a MHC class II gene between steelhead and an upstream isolated resident population of *O. mykiss*. Although not discussed by the authors, a general trend of reduced genetic diversity was observed in the landlocked resident population; however, this population exhibited increased levels of diversity at the MHC-linked marker in accordance with balancing selection within the resident population. More convincing conclusions about balancing selection can be obtained from analyses based on sequencing larger fragments of genes covering multiple polymorphic sites. McCairns et al. (2011) followed this approach for a fragment of the peptide-binding region of a MHC class II gene in stickleback (*Gasterosteus aculeatus*) and found similar evidence for balancing selection. Sequence-based analyses are in general expected to be more powerful for detecting balancing selection compared to individual marker based outlier tests (Renaut et al. 2010; Brieuc and Naish 2011; Narum and Hess 2011).

We also found interleukin genes among outliers in all three genome scans (Fig. 4). A recent study by Narum et al. (2011) also found interesting patterns for these three loci. They found Omy_IL-320 to be a candidate locus for anadromy in *O. mykiss* populations from the Klickitat basin in the Columbia tributary, Washington. This result is in agreement with our observations at this locus showing a signature of divergent selection between the resident and anadromous populations in the Twisp River (Fig. 4D). Furthermore, we observed outlier patterns for two other interleukin markers Omy_IL1b-163 and OmyIL17–185 (Fig. 4B and 4C). These two markers were observed to correlate with one or more environmental variables in the study by Narum et al. (2011), indicative of adaptive roles. For example, Narum et al. (2011) found the Omy_IL1b-163 locus to correlate with temperature, and this finding is also supported here at a larger spatial scale (Table 2). Despite the different spatial scales, our results together with the study by Narum et al. (2011) add strong support for an important adaptive role of interleukin genes in *O. mykiss*. Temperature tolerance, or factors correlating with temperature such as parasite abundance and virulence (e.g., Marcogliese 2008), have also been shown to infer selection on immune genes in other fish and animals in general (e.g., Kurtz et al. 2004; Sommer 2005; McCairns et al. 2011).

In conclusion, we observed interesting patterns of adaptive variation at both interleukin genes (divergent selection) and a MHC class II gene (balancing selection). Here, the latter is represented by three SNPs hitherto unscreened in wild populations. These candidate genes will inevitably prove valuable in future studies of *O. mykiss* investigating the evolutionary role of immune response processes.

The promise of applying functional genetic variation in conservation genomics

Genome scans including functional genetic variation have proven very promising for identifying (and understanding) adaptively important genes and traits in nonmodel organisms (e.g., Namroud et al. 2008; Nielsen et al. 2009b; Glover et al. 2010), also see Vasemägi and Primmer (2005) and Storz (2005) for reviews. First, identification of intraspecific adaptive variation among populations is crucial for identifying focal intraspecific population units of high conservation value. Further, identification of highly discriminatory loci will greatly increase power for use in management related assignment tests (e.g., Freamo et al. 2011) or mixed-stock analyses (Freamo et al. 2011; Seeb et al. 2011b) of natural populations. Future studies identifying adaptive variation are thus expected to contribute toward development of more effective conservation plans at the intraspecific level of wild nonmodel organisms.

Acknowledgments

We thank Carita Pascal (University of Washington) for invaluable laboratory assistance. We also thank Washington Department of Fish and Wildlife (WDFW) Region 4 and Skagit River Cooperative for samples from Puget Sound, Department of Fish and Wildlife Region 6 and US Park Service for samples from the outer coast, and WDFWs Hatchery/Wild Interactions for samples from the Twisp River. Discovery of the SNPs reported in Appendix 2 and development of the assays to interrogate those SNPs was supported by the Washington State General Fund and two Washington State Wildlife Grants. Samples from Canada were kindly provided by Sue Pollard, British Columbia Ministry of Environment. Funding for this project was provided by a grant from the Gordon and Betty Moore Foundation to JES and LWS. MTL received financial support from the European Commission through the FP6 projects UNCOVER (Contract No. 022717) and RECLAIM (Contract No. 044133).

References

Abadia-Cardoso, A., A. J. Clemento, and J. C. Garza. 2011. Discovery and characterization of single-nucleotide polymorphisms in steelhead/rainbow trout, *Oncorhynchus mykiss*. Mol. Ecol. Resour. 11(Suppl 1):31–49.

Aguilar, A., and J. C. Garza. 2006. A comparison of variability and population structure for major histocompatibility complex and microsatellite loci in California coastal steelhead (*Oncorhynchus mykiss* Walbaum). Mol. Ecol. 15:923–937.

Aguilar, A., and J. C. Garza. 2008. Isolation of 15 single nucleotide polymorphisms from coastal steelhead, *Oncorhynchus mykiss* (Salmonidae). Mol. Ecol. Resour. 8:659–662.

Aguilar, A., G. Roemer, S. Debenham, M. Binns, D. Garcelon, and R. K. Wayne. 2004. High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. Proc. Natl. Acad. Sci. U. S. A. 101:3490–3494.

Allendorf, F. W., and F. M. Utter. 1979. Population genetics. Pp. 407–454 *in* W. S. Hoar, D. J. Randall, and J. R. Brett, eds. Fish physiology, Vol. VIII. Academic Press, New York.

Anderson, T. J. C., S. Nair, D. Sudimack, J. T. Williams, M. Mayxay, P. N. Newton, J. P. Guthmann, F. M. Smithuis, T. T. Hien, I. V. F. Van Den Broek, et al. 2005. Geographical distribution of selected and putatively neutral SNPs in Southeast Asian malaria parasites. Mol. Biol. Evol. 22:2362–2374.

Bagley, M. J., and G. A. E. Gall. 1998. Mitochondrial and nuclear DNA sequence variability among populations of rainbow trout (*Oncorhynchus mykiss*). Mol. Ecol. 7:945–961.

Barrett, R. D. H., S. M. Rogers, and D. Schluter. 2008. Natural selection on a major armor gene in threespine stickleback. Science 322:255–257.

Bernatchez, L., and C. Landry. 2003. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J. Evol. Biol. 16:363–377.

Behnke, R. J. 1992. Native trout of Western North America. American Fisheries Society, Monograph 6, Bethesda, MD.

Bierne, N., J. Welch, E. Loire, F. Bonhomme, and P. David. 2011. The coupling hypothesis: why genome scans may fail to map local adaptation genes. Mol. Ecol. 20:2044–2072.

Blankenship, S. M., M. R. Campbell, J. E. Hess, M. A. Hess, T. W. Kassler, C. C. Kozfkay, A. P. Matala, S. R. Narum, M. M. Paquin, M. P. Small, et al. 2011. Major lineages and metapopulations in Columbia River *Oncorhynchus mykiss* are structured by dynamic landscape features and environments. Trans. Am. Fish. Soc. 140:665–684.

Bouck, A., and T. Vision. 2007. The molecular ecologist's guide to expressed sequence tags. Mol. Ecol. 16:907–924.

Brieuc, M. S. O., and K. A. Naish. 2011. Detecting signatures of positive selection in partial sequences generated on a large scale: pitfalls, procedures and resources. Mol. Ecol. Resour. 11:172–183.

Brunelli, J. P., G. H. Thorgaard, R. F. Leary, and J. L. Dunnigan. 2008. Single-nucleotide polymorphisms associated with allozyme differences between inland and coastal rainbow trout. Trans. Am. Fish. Soc. 137:1292–1298.

Campbell, N. R., K. Overturf, and S. R. Narum. 2009. Characterization of 22 novel single nucleotide polymorphism markers in steelhead and rainbow trout. Mol. Ecol. Resour. 9:318–322.

Christie, M. R., M. L. Marine, and M. S. Blouin. 2011. Who are the missing parents? Grandparentage analysis identifies multiple sources of gene flow into a wild population. Mol. Ecol. 20:1263–1276.

Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard. 2010. Using environmental correlations to identify loci underlying local adaptation. Genetics 185:1411–1423.

Currens, K. P., C. B. Schreck, and H. W. Li. 2009. Evolutionary ecology of redband trout. Trans. Am. Fish. Soc. 138:797–817.

Dalziel, A. C., S. M. Rogers, and P. M. Schulte. 2009. Linking genotypes to phenotypes and fitness: how mechanistic biology can inform molecular ecology. Mol. Ecol. 18:4997–5017.

Dijkstra, J. M., I. Kiryu, Y. Yoshiura, A. Kumanovics, M. Kohara, N. Hayashi, and M. Ototake. 2006. Polymorphism of two very similar MHC class Ib loci in rainbow trout (*Oncorhynchus mykiss*). Immunogenetics 58:152–167.

Docker, M. F. and D. D. Heath. 2003. Genetic comparison between sympatric anadromous steelhead and freshwater resident rainbow trout in British Columbia, Canada. Conserv. Genet. 4:227–231.

Elliott, J. 1994. Quantitative ecology and the brown trout. Oxford Univ. Press, Oxford, U.K.

Ewens, W. J. 1972. Sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3:87–112.

Excoffier, L., and H. E. L. Lischer. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10:564–567.

Excoffier, L., T. Hofer, and M. Foll. 2009. Detecting loci under selection in a hierarchically structured population. Heredity 103:285–298.

Foll, M., and O. Gaggiotti. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180:977–993.

Fraser, D. J., L. K. Weir, L. Bernatchez, M. M., Hansen and E. B. Taylor. 2011. Extent and scale of local adaptation in salmonid fishes: review and meta-analysis. Heredity 106:404–420.

Freamo, H., P. O'Reilly, P. R. Berg, S. Lien, and E. G. Boulding. 2011. Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. Mol. Ecol. Resour. 11:254–267.

Glover, K. A., M. M. Hansen, S. Lien, T. D. Als, B. Høyheim, and Ø. Skaala. 2010. A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. BMC Genet. 11:2

Gomez-Uchida, D., J. E. Seeb, M. J. Smith, C. Habicht, T. P. Quinn, and L. W. Seeb. 2011. Single nucleotide polymorphisms unravel hierarchical divergence and signatures of selection among Alaskan sockeye salmon (*Oncorhynchus nerka*) populations. BMC Evol. Biol. 11:48.

Goudet, J. 1995. FSTAT (Version 1.2): a computer program to calculate F-statistics. J. Hered. 86:485–486.

Hansen, J. D., P. Strassburger, G. H. Thorgaard, W. P. Young, and L. Du Pasquier. 1999. Expression, linkage, and polymorphism of MHC-related genes in rainbow trout, *Oncorhynchus mykiss*. J. Immunol. 163:774–786.

Hansen, M. H. H., S. Young, H. B. H. Jorgensen, C. Pascal, M. Henryon, and J. Seeb. 2011. Assembling a dual purpose TaqMan-based panel of single-nucleotide polymorphism markers in rainbow trout and steelhead (*Oncorhynchus mykiss*) for association mapping and population genetics analysis. Mol. Ecol. Resour. 11:67–70.

Harstad, H., M. F. Lukacs, H. G. Bakke, and U. Grimholt. 2008. Multiple expressed MHC class II loci in salmonids; details of one non-classical region in Atlantic salmon (*Salmo salar*). BMC Genomics 9:193.

Heath, D. D., C. M. Bettles, S. Jamieson, I. Stasiak, and M. F. Docker. 2008. Genetic differentiation among sympatric migratory and resident life history forms of rainbow trout in British Columbia. Trans. Am. Fish. Soc. 137:1268–1278.

Helyar, S. J., J. Hemmer-Hansen, D. Bekkevold, M. I. Taylor, R. Ogden, M. T. Limborg, A. Cariani, G. E. Maes, E. Diopere, G. R. Carvalho et al. 2011. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Mol. Ecol. Resour. 11:123–136.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson, and W. A. Cresko. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genetics 6:e1000862.

Jantz, L., R. Kadowaki, and B. Spilsted. 1990. Skeena River salmon test fishery, 1987. *Canadian Data Report of Fisheries and Aquatic Sciences, No. 804*. Prince Rupert, BC.

Kurtz, J., M. Kalbe, P. B. Aeschlimann, M. A. Haberli, K. M. Wegner, T. B. H. Reusch, and M. Milinski. 2004. Major histocompatibility complex diversity influences parasite resistance and innate immunity in sticklebacks. Proc. R. Soc. Lond. Ser. B 271:197–204.

Landry, C., and L. Bernatchez. 2001. Comparative analysis of population structure across environments and geographical scales at major histocompatibility complex and microsatellite loci in Atlantic salmon (*Salmo salar*). Mol. Ecol. 10:2525–2539.

Lewontin, R. C. 1970. The genetic basis of evolutionary change. Columbia Univ. Press, New York.

MacCrimmon, H. R. 1971. World distribution of rainbow trout (*Salmo gairdneri*). J. Fish. Res. Board Can. 28:663–704.

Marcogliese, D. J. 2008. The impact of climate change on the parasites and infectious diseases of aquatic animals. Rev. Sci. Tech. OIE 27:467–484.

Martínez, A., J. C. Garza, and D. E. Pearse. 2011. A microsatellite genome screen identifies chromosomal regions under differential selection in steelhead and rainbow trout (*Oncorhynchus mykiss*). Trans. Am. Fish. Soc. 140:829–842.

McCairns, R. J. S., S. Bourget, and L. Bernatchez. 2011. Putative causes and consequences of MHC variation within and between locally adapted stickleback demes. Mol. Ecol. 20:486–502.

McCusker, M. R., E. Parkinson, and E. B. Taylor. 2000. Mitochondrial DNA variation in rainbow trout (*Oncorhynchus mykiss*) across its native range: testing biogeographical hypotheses and their relevance to conservation. Mol. Ecol. 9:2089–2108.

McGlauflin, M. T., D. Schindler, C. Habicht, L. W. Seeb, and J. E. Seeb. 2011. Influences of spawning habitat and geography: population structure and juvenile migration timing of sockeye salmon in the Wood River lakes, Alaska. Trans. Am. Fish. Soc. 140:763–782.

Meier, K., M. M. Hansen, D. Bekkevold, O. Skaala, and K. L. D. Mensberg. 2011. An assessment of the spatial scale of local adaptation in brown trout (*Salmo trutta* L.): footprints of selection at microsatellite DNA loci. Heredity 106:488–499.

Miller, K. M., K. H. Kaukinen, T. D. Beacham, and R. E. Whitler. 2001. Geographic heterogeneity in natural selection on an MHC locus in sockeye salmon. Genetica 111:237–257.

Miller, K. M., S. Li, T. J. Ming, K. H. Kaukinen, and A. D. Schulze. 2006. The salmonid MHC class I: more ancient loci uncovered. Immunogenetics 58:571–589.

Morin, P. A., G. Luikart, and R. K. Wayne. 2004. SNPs in ecology, evolution and conservation. Trends Ecol. Evol. 19:208–216.

Namroud, M. C., J. Beaulieu, N. Juge, J. Laroche, and J. Bousquet. 2008. Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. Mol. Ecol. 17:3599–3613.

Narum, S. R., and J. E. Hess. 2011. Comparison of F-ST outlier tests for SNP loci under selection. Mol. Ecol. Resour. 11:184–194.

Narum, S. R., J. S. Zendt, D. Graves, and W. R. Sharp. 2008. Influence of landscape on resident and anadromous life history types of *Oncorhynchus mykiss*. Can. J. Fish. Aquat. Sci. 65:1013–1023.

Narum, S. R., N. Campbell, A. Matala, and J. Hess. 2010a. Genetic assessment of Columbia River stocks. Columbia River Inter-Tribal Fish Comission, Technical Report 10–12. Portland, OR.

Narum, S. R., N. R. Campbell, C. C. Kozfkay, and K. A. Meyer. 2010b. Adaptation of redband trout in desert and montane environments. Mol. Ecol. 19:4622–4637.

Narum, S. R., J. S. Zendt, C. Frederiksen, N. Campbell, A. Matala, and W. R. Sharp. 2011. Candidate genetic markers associated with anadromy in *Oncorhynchus mykiss* of the Klickitat River. Trans. Am. Fish. Soc. 140:843–854.

Nielsen, E. E., J. Hemmer-Hansen, P. F. Larsen, and D. Bekkevold. 2009a. Population genomics of marine fishes: identifying adaptive variation in space and time. Mol. Ecol. 18:3128–3150.

Nielsen, E. E., J. Hemmer-Hansen, N. A. Poulsen, V. Loeschcke, T. Moen, T. Johansen, C. Mittelholzer, G. L. Taranger, R. Ogden, and G. R. Carvalho. 2009b. Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). BMC Evol. Biol. 9:276.

Paschou, P., E. Ziv, E. G. Burchard, S. Choudry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas. 2007. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genetics 3:1672–1686.

Peakall, R., and P. E. Smouse. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol. Ecol. Notes 6:288–295.

Pearse, D. E., S. A. Hayes, M. H. Bond, C. V. Hanson, E. C. Anderson, R. B. Macfarlane, and J. C. Garza. 2009. Over the Falls? Rapid Evolution of Ecotypic Differentiation in Steelhead/Rainbow Trout (*Oncorhynchus mykiss*). J. Hered. 100:515–525.

Piertney, S. B., and M. K. Olivier. 2006. The evolutionary ecology of the major histocompatibility complex. Heredity 96:7–21.

Quinn, T. P. 2005. The behaviour and ecology of Pacific salmon and trout. American Fisheries Society/Univ. of Washington Press, BethesdaMD / SeattleWA.

Renaut, S., A. W. Nolte, and L. Bernatchez. 2010. Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus spp.* Salmonidae). Mol. Ecol. 19:115–131.

Rice, W. R. 1989. Analyzing tables of statistical tests. Evolution 43:223–225.

Rousset, F. 2008. GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. Mol. Ecol. Resour. 8:103–106.

Rubidge, E. M., and E. B. Taylor. 2004. Hybrid zone structure and the potential role of selection in hybridizing populations of native westslope cutthroat trout (*Oncorhynchus clarki lewisi*) and introduced rainbow trout (*O. mykiss*) . Mol. Ecol. 13:3735–3749.

Sanchez, C., T. Smith, R. Wiedmann, R. Vallejo, M. Salem, J. Yao, and C. Rexroad. 2009. Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. BMC Genomics 10:559.

Schlötterer, C. 2004. The evolution of molecular markers – Just a matter of fashion? Nat. Rev. Genet. 5:63–69.

Seeb, J. E., C. E. Pascal, R. Ramakrishnan, and L. W. Seeb. 2009. SNP genotyping by the 5'-nuclease reaction: advances in high throughput genotyping with non-model organisms. Pp. 277–292 *in* A. Komar, ed. Methods in molecular biology, single nucleotide polymorphisms. 2nd ed. Humana Press, New York.

Seeb, J. E., G. Carvalho, L. Hauser, S. Roberts, L. W. Seeb, and K. Naish. 2011a. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Mol. Ecol. Resour. 11:1–8.

Seeb, L. W., W. D. Templin, S. Sato, S. Abe, K. Warheit, J. Y. Park, and J. E. Seeb. 2011b. Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. Mol. Ecol. Resour. 11:195–217.

Sommer, S. 2005. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. Front. Zool. 2:16.

Stephens, M., N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68:978–989.

Stephens, M. R., N. W. Clipperton, and B. May. 2009. Subspecies-informative SNP assays for evaluating introgression between native golden trout and introduced rainbow trout. Mol. Ecol. Resour. 9:339–343.

Storz, J. F. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol. Ecol. 14:671–688.

Taylor, E. B., C. J. Foote, and C. C. Wood. 1996. Molecular genetic evidence for parallel life-history evolution within a Pacific salmon (sockeye salmon and kokanee, *Oncorhynchus nerka*). Evolution 50:401–416.

Utter, F. 2004. Population genetics, conservation and evolution in salmonids and other widely cultured fishes: some perspectives over six decades. Rev. Fish Biol. Fisher. 14:125–144.

Utter, F., D. Campton, S. Grant, G. Milner, J. Seeb, and L. Wishard. 1980. Population structures of indigenous salmonid species of the Pacific Northwest. Pp. 285–304 *in* W. J. McNeil, and D. C. Himsworth, eds. Salmonid ecosystems of the North Pacific. Oregon State Univ., Corvalis.

Vasemägi A, and C. R. Primmer. 2005. Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. Mol. Ecol. 14:3623–3642.

Waples, R. S. 1990. Temporal changes of allele frequency in Pacific Salmon – implications for mixed-stock fishery analysis. Can. J. Fish. Aquat. Sci. 47:968–976.

Watterson, G. A. 1978. Homozygosity test of neutrality. Genetics 88:405–417.

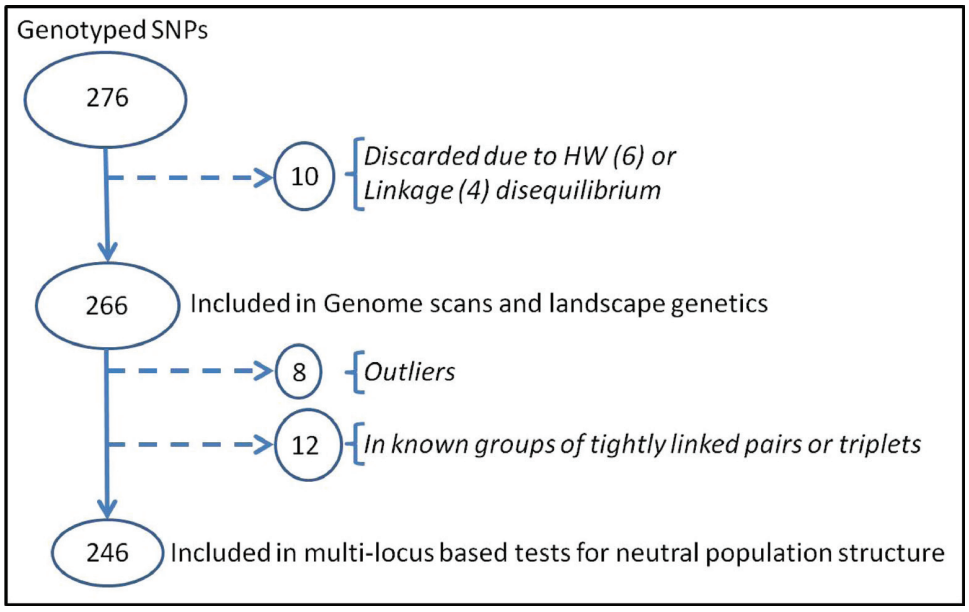
Wenger, S. J., D. J. Isaak, C. H. Luce, H. M. Neville, K. D. Fausch, J. B. Dunham, D. C. Dauwalter, M. K. Young, M. M. Elsner, B. E. Rieman, et al. 2011. Flow regime, temperature, and biotic interactions drive differential declines of trout species under climate change. Proc. Natl. Acad. Sci. U. S. A. 108:14175–14180.

Wishard, L. N., J. E. Seeb, F. M. Utter, and D. Stefan. 1984. A genetic investigation of suspected redband trout populations. Copeia 1984:120–132.

Young, F. 1996. ViSta: the visual statistics system. 2nd ed. Research Memorandum 94-I (b). L.L. Thursone Psychometric Laboratory, Univ. of North Carolina, Chaper Hill, NC.

Zimmerman, C. E. and G. H. Reeves. 2000. Population structure of sympatric anadromous and nonanadromous *Oncorhynchus mykiss*: evidence from spawning surveys and otolith microchemistry. Can. J. Fish. Aquat. Sci. 57:2152–2162.

Appendix



Appendix 1. SNP exclusion pipeline. Ellipses isolate numbers of SNPs in the various categories. Solid arrows connect the stages of analysis, and broken arrows identify sets of SNPs that were excluded at each stage.

Appendix 2. TaqMan assays for *O. mykiss* developed in the Washington Department of Fish and Wildlife (WDFW) Molecular Genetics Laboratory for genotyping steelhead. Gene targets were identified by BLASTX alignments in the Swiss-Prot database with E-values <1 and amino acid identities >25%.

Assay name	Gene target	Design strand	Primers	Dye label	Probe sequences	Allele calls	Ascertainment set*
Omy_BAC-F5-284	Unknown	F	ACAACGCCAACAACTTTCTCTTG	VIC	CAGTAGGGCGGCAAG	C	1
		R	CCTCATTTACTGTAGGACCATGCA	FAM	ACAGTAGGACGGCAAG	T	
Omy_Il-1b_-028	Interleukin 1 beta_2	F	ACTGTCTGGCTAGAGCACATTG	VIC	CTGAGGCAACTTTTGT	T	2
		R	ATCTTCTACCACCGCACTGTTTTAA	FAM	TGAGGCAGCTTTTGT	C	
Omy_UT11_2a-018	Clock homolog	F	GCCTCCTGTGTATGAGTATGTTAAGTT	VIC	CTGAATAATGGTAAGTTTTC	T	2
		R	GCCAGCTACTTGTGTACTACATTCA	FAM	TGAATAATGGTACGTTTTC	G	
Omy_UT11_2a-114	Clock homolog	F	GCTGGCCTGTATTTGGTTTGTG	VIC	TGGTCGGTCTAAGTGC	C	2
		R	TTGGAGTTGAGTTAGGCACTATAGGA	FAM	TGGTCAGTCTAAGTGC	T	
Omy_UT11_2a-132	Clock homolog	F	GCTGGCCTGTATTTGGTTTGTG	VIC	AGGCAACAATGACTAATT	C	2
		R	CCATTTGGAGTTGAGTTAGGCACTA	FAM	AAGGCAACAATTACTAATT	A	
Omy_u07-79.166	Unknown	F	CCCGCTATATTATTTGATCACCCTTGA	VIC	ACTTGGGAATACCCCAGCC	G	1
		R	ATTTAAATCCATTTCTAAAAATAAGCAAACCTAACCA	FAM	CTTGGGAATAACCCAGCC	T	
Omy_u09-52.284	Unknown	F	TTTGTGTGTATTGTTGTGACTTG	VIC	ACTGCATTGTTGTAGCTAG	T	1
		R	TGATGTTATTGCAGGTCTAGCGAAA	FAM	CTGCATTGTTGTCGCTAG	G	
Omy_u09-53.469	Unknown	F	ACAGCCTGAGCGTTTGCA	VIC	TTGCAGCCCTTATTGTG	T	1
		R	GGAAACTGGGAGAGATCAAAGGA	FAM	TTGCAGCCCTTGTTGTG	C	
Omy_09-55-233	Unknown	F	CTCGTTTGATAGAAGAAACAAAGTGAAAGTG	VIC	AGCACTGACATCTGC	A	1
		R	CCAACATCTTTGGGCCTAAACAAGA	FAM	AGCACTGACATCCGC	G	
Omy_u09-56.073	Unknown	F	CCCACTACATCCTCATCCAAGGT	VIC	AGCGGCATTCTC	C	1
		R	CTCACTGCAAATCCAACTTCATCAT	FAM	CAGCGTCATTCTC	A	
Omy_u09-56.119	Unknown	F	CCAAGGTGGACCCACCAG	VIC	AGTGAGCTGAAACAGAGCA	T	1
		R	GCTGAGTTTATAGGTCAGTCATTATACATATTGA	FAM	TGAGCTGAAGCAGAGCA	C	
Omy_M09AAC-301	Polymerase (DNA directed), epsilon 2 (p59 subunit)	F	CATTGATGGTTATTGTCATGCGTTTCA	VIC	CTGACAGATTTTGAAGTCT	T	1
		R	GCAGTAGAGATAGAAATTGAGCACACT	FAM	CTGACAGATTTTGACGTCT	G	
Omy_M09AAF-098	Exostosin like	F	CGGCGCCCGTCAGTA	VIC	CACAGGAGTAGTTAGAGTTA	T	1
		R	CAACCACCAGTCATGGCTCTA	FAM	CACAGGAGTAGTTCGAGTTA	G	
Omy_M09AAD-076	Family with sequence similarity 203, member A	F	ACTGTTACCACTCTCTCATCAACCT	VIC	CACCAACCACTGGTGAA	T	1
		R	GGGTCCAGGAGGTTTTTAAACAACAT	FAM	CCAACCGCTGGTGAA	C	
Omy_09AAH-172	Rab-like protein 2A	F	GGCGTGAGCTTTGTGGTAGTA	VIC	AGGGATGCATACTCTG	T	1
		R	GCTGGACAGGAGCGGTTC	FAM	AGGGATGCATGCTCTG	C	
Omy_09AAI-092	Ubiquitin-like modifier activating enzyme 3	F	GGGCCTTGCTTTGCTTCTG	VIC	ATCTACTGAGTCTCTGCTGC	G	1
		R	ACGGCTTAGATCCTCCTTGGA	FAM	CTACTGAGTGTCTGCTGC	C	

*Ascertainment set 1 included steelhead from Kamchatka, Russia ($n = 3$), Alaska, USA ($n = 3$), British Columbia, Canada ($n = 3$), Washington, USA ($n = 180$). Ascertainment set 2 included rainbow trout from Kamchatka, Russia ($n = 10$), Alaska, USA ($n = 10$), Washington, USA ($n = 70$), and cutthroat trout from Montana, USA ($n = 10$).

Appendix 3. Tested environmental variables and data sources.

Environmental variable	SUST[1]	SKAG	GREEN	SOL	CHEH	DESCH	CROOK	TWISP	DEADM[1]
Precipitation (mm)[2]	965	873	1082	2501	1413	279	979	395	263
Maximum temperature (°C)[3]	21.0	22.7	24.6	20.0	24.8	32.2	28.0	29.5	26.2
Minimum temperature (°C) [4]	–13.3	0.9	1.7	1.9	1.0	–3.5	–8.7	–9.6	–9.2
Elevation (m) [5]	1352	9	22	9	28	397	1049	494	336
Latitude	56.58	48.44	47.29	47.91	46.80	44.82	46.51	48.37	50.74
Longitude	–126.45	–122.34	–122.17	–124.54	–123.17	–121.09	–114.68	–120.14	–120.92

Sources:

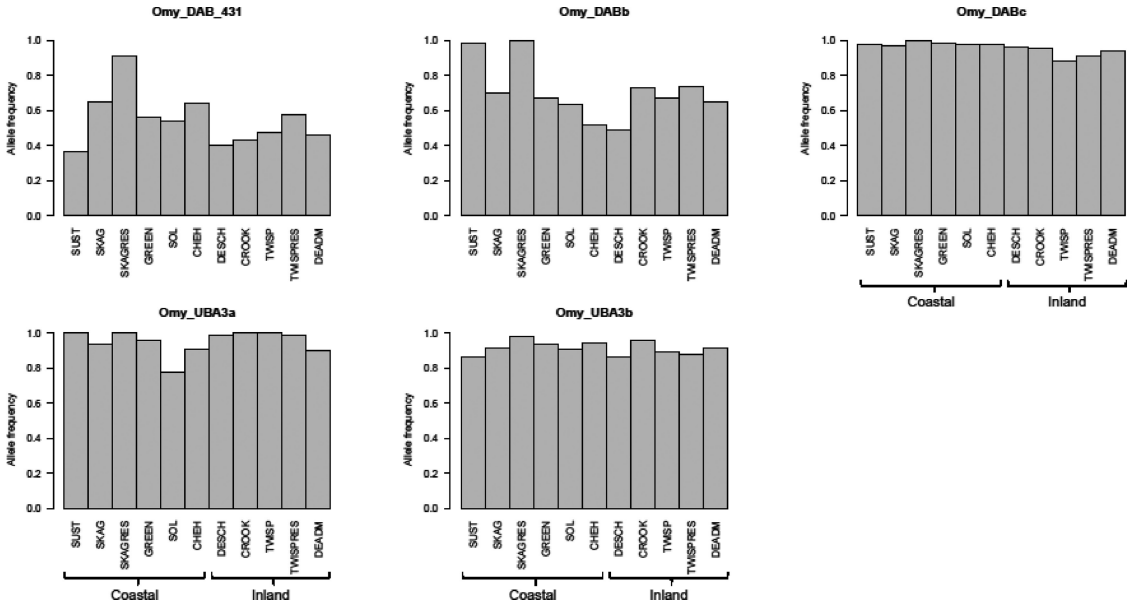
[1]Precipitation and temperature data for the BC samples obtained from http://www.genetics.forestry.ubc.ca/cfcg/climate-models.html

[2]http://www.prism.oregonstate.edu/products/matrix.phtml?vartype=ppt&view=maps.

[3]http://www.prism.oregonstate.edu/products/matrix.phtml?vartype=tmax&view=maps.

[4]http://www.prism.oregonstate.edu/products/matrix.phtml?vartype=tmin&view=maps.

[5]Obtained from Google Earth using coordinates.



Appendix 4. Frequency plots of major allele frequencies for five MHC-related SNPs. Omy_DAB_431 and Omy_DABb are nonsynonymous mutations in the peptide-binding region of a MHC class II gene, while Omy_DABc represents a synonymous mutation in the same gene. Omy_UBA3a and Omy_UBA3b show allele frequencies for two SNPs residing within a MHC class I gene.

Supporting Information

Additional Supporting Information may be found online on Wiley Online Library.

Table S1. Locus information and summary statistics

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Chapter 9

Future directions and perspectives

Future directions and perspectives

In this thesis, I defined concrete objectives aiming to increase our understanding about the evolutionary mechanisms underlying population structure and local adaptation in marine and salmonid fishes. These questions were successfully addressed using population genetics and genomics approaches, but during the course of my work new exciting questions arose along the way, some of which have been discussed in *chapter 1*. In the following I will briefly put these ideas into perspective in a discussion of where I believe the future of ecological genomics will take us on the on-going quest of increasing our knowledge about natural selection in the wild.

With the continuing increase in speed and reduced costs for performing large scale sequencing analyses, it will eventually be feasible to sequence full genomes even in population level studies. However, such approaches will probably not be worth the effort since a dense genome-wide marker set should be adequate for capturing overall genomic variation due to linkage disequilibrium (Allendorf et al. 2010). Once a reference genome (or linkage map) has been generated for a given species, it can be used to tackle a range of specific questions using marker based approaches including the generation of genome wide marker panels (Hansen et al. 2011), reduced representation sequencing (Sanchez et al. 2009), as well as mapping and identification of candidate genes (e.g. Bradbury et al. 2010; Hohenlohe et al. 2010).

The challenge of distinguishing past from on-going selection should receive more attention as this may have implications for forecasting evolutionary responses of ongoing climate changes. One way to approach this question is to use dense marker coverage around selected loci for inferring the background of adaptive mutations by distinguishing between old sweeps or more contemporary selection (Bierne 2010). Two of the species studied in this PhD represent excellent research models for addressing this question. The inland and coastal *Oncorhynchus mykiss* lineages represent a good model system for assessing such varying temporal scales of selection. Observed signatures may either represent imprints from glacial selection between differing refugial environments or more recent selection imposed by contemporary habitats (see *chapter 8*). Indeed, the necessary genomic resources for such inferences are expected to be available for *O. mykiss* in the near future (Miller et al. 2011). Another interesting case is the different herring populations inhabiting isolated low salinity environments in the Baltic Sea and Ringkøbing fjord, where at least the latter is relatively newly established (<400 years). It would hence be of interest to establish if salinity related adaptive signatures were caused *in* or *ex situ* (*chapter 7*). Thus, observed patterns of divergent selection in relation to low salinity in this area

are likely to, at least partly, represent historical adaptation in an ancestral population. It is likely though, that on-going selection on adaptive alleles for low salinity maintains divergent allele frequencies between neighbouring marine and brackish populations, and that both historical and contemporary selection explains the observed adaptive patterns.

Improved temporal understanding will also be crucial for optimising inferences gained through landscape genomics approaches, which will undoubtedly improve our knowledge about the interacting effects of gene-flow, genetic drift and natural selection in varying and changing environments. Being able to distinguish neutral from adaptive variation and to identify environmental boundaries impeding gene-flow appears extremely promising for the detection of demographically independent units which should serve as conservation units (Ouborg et al. 2010). Further, increased understanding of adaptive genes and associated key environmental variables can be used to predict future population dynamics by modelling gene-flow and adaptation in forecasted landscapes under various climate change scenarios (Allendorf et al. 2010). For example, Wenger et al. (2011) predicted drastic declines in suitable habitat for four trout species by considering forecasted changes of key environmental factors demonstrating the usefulness of such an approach.

Despite the intriguing accomplishments waiting around the corner from fully sequenced and annotated genomes, these will not close the existing knowledge gap between genetic variation and ultimate performance of organisms in the wild (Figure 1). While increased genomic inference may serve as a useful knowledge base for identifying candidate genes and key environmental factors, inference from other biological processes is required to fully understand the interplay between the genetic and environmental drivers shaping the phenotype. Thus, direct inference from higher biological levels is urgently needed. One promising approach to link genetic variation with “quantitative” phenotypic traits in fishes is to use QTL mapping (Figure 1; Naish and Hard 2008). Currently very few applications in wild nonmodel fishes exist (but see Rogers and Bernatchez 2007), however, this number is expected to increase in the near future reflecting the fast accumulation of comprehensive genomic resources in many nonmodel fishes. Furthermore, inferences from the transcriptome and proteome (Figure 1) appear especially promising for understanding higher level biological processes shaping phenotypes in wild organisms (Dalziel et al. 2009). Although underlying theory and methods in the fields of transcriptomics and proteomics are not as developed as in genetics (Anderson and Anderson 1998), future advances within these fields will be able to illuminate new sides of the genotype–

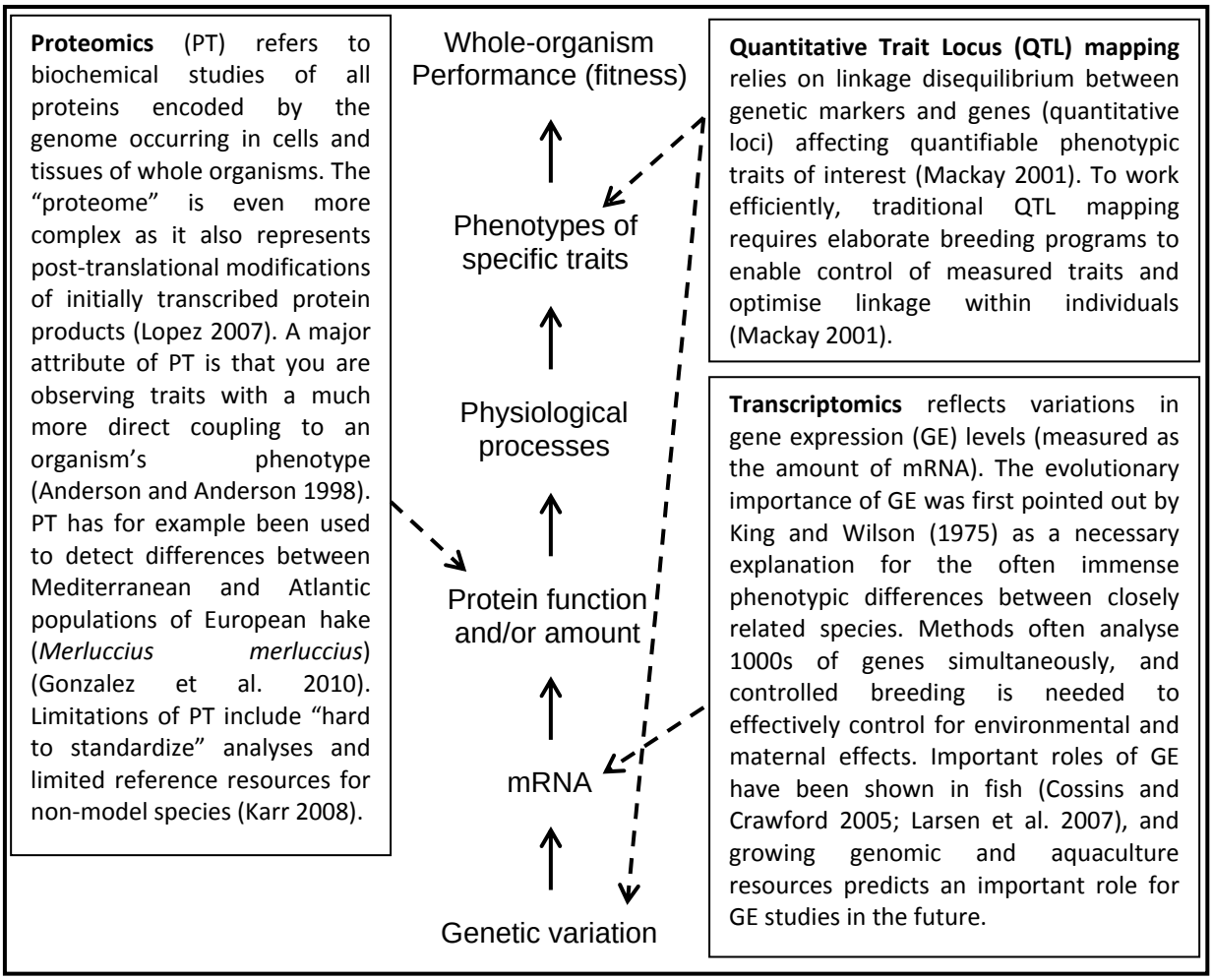


Figure 1 The genotype-phenotype pathway revisited from figure 1 in *chapter 1*. Boxes describe how proteomic, QTL and transcriptomic approaches draw inference from different biological levels.

phenotype pathway by elucidating a wider range of the "biological functions" allowing populations to adapt to new environmental conditions (Pandey and Mann 2000).

In conclusion, it appears obvious that much is to be gained from closer collaborations among disciplines. Research at genetic and physiological levels, which conceptually address identical questions, should be increasingly united within scopes of scientific journals and conferences. Whereas Bob Dylan's famous words "*The Times They Are a-Changin*" may seem universally fitting, the rate by which times are changing has never been faster within this field, and the future does indeed look promising – We'd better hold on.

References

Allendorf, F. W., P. A. Hohenlohe and G. Luikart (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11**: 697-709.

Anderson, N. L. and N. G. Anderson (1998). Proteome and proteomics: New technologies, new concepts, and new words. *Electrophoresis* **19**: 1853-1861.

Bierne, N. (2010). The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution* **64**: 3254-3272.

Bradbury, I. R., S. Hubert, B. Higgins, T. Borza, S. Bowman, I. G. Paterson, P. V. R. Snelgrove, C. J. Morris, R. S. Gregory, D. C. Hardie, et al. (2010). Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society B-Biological Sciences* **277**: 3725-3734.

Cossins, A. R. and D. L. Crawford (2005). Opinion - Fish as models for environmental genomics. *Nature Reviews Genetics* **6**: 324-333.

Dalziel, A. C., S. M. Rogers and P. M. Schulte (2009). Linking genotypes to phenotypes and fitness: how mechanistic biology can inform molecular ecology. *Molecular Ecology* **18**: 4997-5017.

Gonzalez, E. G., G. Krey, M. Espineira, A. Diez, A. Puyet and J. M. Bautista (2010). Population Proteomics of the European Hake (*Merluccius merluccius*). *Journal of Proteome Research* **9**: 6392-6404.

Hansen, M. H. H., S. Young, H. B. H. Jorgensen, C. Pascal, M. Henryon and J. Seeb (2011). Assembling a dual purpose TaqMan-based panel of single-nucleotide polymorphism markers in rainbow trout and steelhead (*Oncorhynchus mykiss*) for association mapping and population genetics analysis. *Molecular Ecology Resources* **11**: 67-70.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson and W. A. Cresko (2010). Population genomics of parallel adaptation in Threespine stickleback using sequenced RAD Tags. *PLoS Genetics* **6**: e1000862.

Karr, T. L. (2008). Application of proteomics to ecology and population biology. *Heredity* **100**: 200-206.

King, M. C. and A. C. Wilson (1975). Evolution at 2 Levels in Humans and Chimpanzees. *Science* **188**: 107-116.

Larsen, P. F., E. E. Nielsen, T. D. Williams, J. Hemmer-Hansen, J. K. Chipman, M. Kruhoffer, P. Grønkjær, S. G. George, L. Dyrskjot and V. Loeschcke (2007). Adaptive differences in gene expression in European flounder (*Platichthys flesus*). *Molecular Ecology* **16**: 4674-4683.

Lopez, J. L. (2007). Applications of proteomics in marine ecology. *Marine Ecology-Progress Series* **332**: 275-279.

Mackay, T. F. C. (2001). The genetic architecture of quantitative traits. *Annual Review of Genetics* **35**: 303-339.

Miller, M. R., J. P. Brunelli, P. A. Wheeler, S. Liu, C. E. Rexroad, Y. Palti, C. Q. Doe and G. H. Thorgaard (2011). A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology* **doi: 10.1111/j.1365-294X.2011.05305.x**.

Naish, K. A. and J. J. Hard (2008). Bridging the gap between the genotype and the phenotype: linking genetic variation, selection and adaptation in fishes. *Fish and Fisheries* **9**: 396-422.

Ouborg, N. J., C. Pertoldi, V. Loeschcke, R. Bijlsma and P. W. Hedrick (2010). Conservation genetics in transition to conservation genomics. *Trends in Genetics* **26**: 177-187.

Pandey, A. and M. Mann (2000). Proteomics to study genes and genomes. *Nature* **405**: 837-846.

Rogers, S. M. and L. Bernatchez (2007). The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (*Coregonas sp* Salmonidae) species pairs. *Molecular Biology and Evolution* **24**: 1423-1438.

Sanchez, C., T. Smith, R. Wiedmann, R. Vallejo, M. Salem, J. Yao and C. Rexroad (2009). Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics* **10**: 559.

Wenger, S. J., D. J. Isaak, C. H. Luce, H. M. Neville, K. D. Fausch, J. B. Dunham, D. C. Dauwalter, M. K. Young, M. M. Elsner, B. E. Rieman, et al. (2011). Flow regime, temperature, and biotic interactions drive differential declines of trout species under climate change. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 14175-14180.